

BASIC PHYSICS

PRINCIPLES AND CONCEPTS

Avijit Lahiri



BASIC PHYSICS

PRINCIPLES AND CONCEPTS

Avijit Lahiri

Basic Physics: Principles and Concepts

(an e-book)

All rights reserved.

by **Avijit Lahiri**, author and publisher.

252 Laketown, Block A, Kolkata 700089, India.

e-mail: avijit.lahiri.al@gmail.com / website: physicsandmore.net

March 2012 / November 2018.

[Revision] March, 2020

**To my parents
whom I miss**

To Anita, Anindita, Indrajit

**And to my students
who gave me much of my basic physics
while being taught by me**

ACKNOWLEDGEMENT

First of all, I acknowledge huge support from Pradip Chatterjee, who never failed to sit through discussion sessions, helping me clarify my ideas.

Sankhasubhra Nag and Saugata Bhattacharyya were there whenever I needed them all through the writing of this book.

Pariksheet Manna and Basanta Halder smilingly offered technical help in preparing the manuscript.

Above all, it is a pleasure to acknowledge great emotional support from my wife Anita, and from children Anindita and Indrajit.

Whatever errors and oddities remain in the end are, however, solely my own doing.

Kolkata, March, 2012 / November, 2018.

Revision: March, 2020

List of chapters

1. Introduction
2. Vectors
3. Mechanics
4. Simple Harmonic Motion
5. Gravitation
6. Elasticity
7. Mechanics of fluids
8. Thermal Physics
9. Wave motion I: Acoustic waves
10. Ray Optics
11. Electrostatics
12. Electricity I: Steady currents and their magnetic effects
13. Electricity II: Varying and alternating currents
14. Wave motion II: Electromagnetic waves
15. Wave Optics
16. Quantum theory
17. Relativity: the special and the general theory
18. Atoms, Nuclei, and Molecules
19. Electronics

Contents

1	Introduction: Units and Dimensions	1
1.1	Physical quantities and their units	1
1.2	Systems of units: the SI system	2
1.2.1	Relations among physical quantities, and their units	2
1.2.2	The dimension of a physical quantity	3
1.3	Basic and derived units	4
1.4	SI units, and dimensions	6
1.4.1	The seven base units	6
1.4.2	Dimensions related to units	8
1.4.3	Derived units: selected physical quantities	10
1.4.4	Units and dimensions of a few physical constants	15
1.4.5	Prefixes denoting multiples and submultiples	18
1.5	Other systems of units	18
1.5.1	Systems of units other than the SI system	18
1.5.2	Conversion from the SI to other systems of units	19
1.5.3	A few convenient non-SI units	19
1.6	Dimensional analysis	21
1.6.1	Principle of dimensional homogeneity	21
1.6.2	An application: Stokes' formula for viscous drag force	22
1.6.3	The principle of similarity	24
1.7	Physical quantities as scalars and vectors	27
2	Vectors	31
2.1	Introduction	31
2.1.1	Equality of two vectors	33

CONTENTS

2.1.2	Magnitude of a vector	34
2.1.3	The null vector	35
2.2	Operations with vectors	35
2.2.1	Addition of vectors	35
2.2.1.1	Addition of two vectors	35
2.2.1.2	Addition of more than two vectors	36
2.2.2	Multiplication of a vector with a scalar	37
2.2.3	Features of vector addition and scalar multiplication	38
2.3	Unit vector	40
2.4	Scalar product of two vectors	42
2.4.1	Features of scalar product	43
2.4.2	Orthonormal triads of vectors	44
2.5	Cartesian components of a vector	45
2.5.1	Two dimensional vectors	47
2.5.2	Vector operations in terms of Cartesian components	48
2.5.3	Scalar product of two vectors in terms of Cartesian components	49
2.5.4	Direction cosines relating to a unit vector	50
2.6	The vector product of two vectors	52
2.6.1	Features of the vector product	53
2.6.2	Vector expression for a planar area	55
2.7	Scalar and vector components of a vector	56
2.8	The right hand rule	59
2.9	Transformation of vectors	60
2.10	Scalar and vector triple products	64
2.10.1	Scalar triple product	64
2.10.2	Vector triple product	66
2.11	Vector function of a scalar variable	67
2.11.1	The derivative of a vector function	68
2.12	Scalar function of a vector variable	70
2.12.1	The position vector	70
2.12.2	Scalar function of the position vector: scalar field	72

CONTENTS

2.13	Vector function of a vector variable: vector field	73
2.14	Derivatives and integrals of scalar and vector fields	74
2.14.1	Derivatives of scalar and vector fields	74
2.14.2	Volume-, surface-, and line integrals	76
3	Mechanics	81
3.1	Introduction: frames of reference	81
3.2	Motion of a single particle	84
3.2.1	Introduction	84
3.2.2	Kinematic quantities	86
3.2.2.1	The position vector and its time dependence	87
3.2.2.2	Displacement	88
3.2.2.3	Velocity	89
3.2.2.4	Speed	90
3.2.2.5	Planar and rectilinear motions	91
3.2.2.6	Momentum	93
3.2.2.7	Kinetic energy	94
3.2.2.8	Acceleration	95
3.3	Newton's laws of motion: Introduction	96
3.4	Inertial frames. Newton's first law.	97
3.5	Newton's second law: Equation of motion	98
3.5.1	The concept of force	98
3.5.2	The second law	99
3.5.3	Equation of motion	100
3.5.4	The line of action of a force	102
3.5.5	The resultant of a number of forces	104
3.5.6	Forces in equilibrium	105
3.6	Motion along a straight line	106
3.7	Motion in a plane	111
3.8	Field of force	116
3.9	Transformations from one frame of reference to another	117
3.9.1	Transformation of velocity	117

CONTENTS

3.9.2	Relative velocity	119
3.9.3	Transformations of displacement and time	121
3.9.4	Transformation of acceleration	123
3.10	Equations of motion in different frames of reference	124
3.10.1	Characterization of the inertial frames	124
3.10.2	Equations of motion of a particle in inertial frames	124
3.10.3	Equation of motion in a non-inertial frame	126
3.11	Force and work	130
3.12	Work in rectilinear motion. Potential energy.	134
3.13	Kinetic energy	137
3.14	Principle of conservation of energy in rectilinear motion	139
3.15	Kinetic energy, potential energy, and work	140
3.15.1	Kinetic energy and work	140
3.15.2	Potential energy. Conservative force-fields.	141
3.15.3	Principle of conservation of energy: the general context	144
3.15.4	Work and energy: summary	145
3.15.5	Energy as ability to perform work	148
3.15.6	Conservation of energy: a broader view	150
3.16	Mechanics of a single particle: overview	154
3.17	Mechanics of a system of particles	156
3.17.1	Internal and external forces. Equations of motion.	156
3.17.2	Newton's third law	157
3.17.3	Center of mass of a system of particles	159
3.17.3.1	The position and velocity of the center of mass	159
3.17.3.2	center of mass momentum	162
3.17.3.3	Determination of the center of mass	163
3.17.4	System of particles: center of mass motion	165
3.17.5	Principle of conservation of momentum	167
3.17.6	System of particles: principle of conservation of energy	168
3.17.7	Elastic and inelastic collisions	170
3.17.7.1	Introduction	170

CONTENTS

3.17.7.2	Elastic and inelastic processes	171
3.17.7.3	The energy balance equation	172
3.17.7.4	Momentum balance	174
3.17.7.5	Relative velocities: normal and tangential	174
3.17.7.6	The center of mass frame	176
3.17.7.7	Describing the collision process	176
3.17.7.8	'Head-on' collision	177
3.17.7.9	Head-on collision in the 'laboratory frame'	179
3.17.7.10	Elastic collisions in planar motion	181
3.17.7.11	Direction of energy transfer in elastic collisions	185
3.17.8	Impulse. Impulsive forces.	188
3.17.8.1	Impulse of a force or of a system of forces	188
3.17.8.2	Impulsive forces	191
3.18	Newton's laws and action-at-a-distance	194
3.19	Angular motion	196
3.19.1	Angular velocity of a particle about a point	196
3.19.2	Angular velocity about an axis	199
3.19.3	Circular motion	201
3.19.4	Centripetal acceleration	204
3.19.4.1	Uniform circular motion	204
3.19.4.2	Motion along a space curve	209
3.19.5	Radial and cross-radial accelerations in planar motion	211
3.19.6	Angular momentum	214
3.19.6.1	Angular momentum about a point	214
3.19.6.2	Angular momentum about an axis	215
3.19.6.3	Angular momentum in circular motion	216
3.19.7	Moment of a force	217
3.19.7.1	Moment of a force about a point	217
3.19.7.2	Moment of a force about an axis	220
3.19.7.3	Impulse of a torque	221
3.19.8	Angular motion of a system of particles	222

CONTENTS

3.19.9	Principle of conservation of angular momentum	223
3.19.10	Rotational motion about an axis: moment of inertia	225
3.19.11	Work done in rotational motion	231
3.19.12	Potential energy in rotational motion	233
3.19.13	Calculation of moments of inertia	234
3.19.13.1	The theorem of perpendicular axes	234
3.19.13.2	The theorem of parallel axes	236
3.19.13.3	Radius of gyration	237
3.19.13.4	Additivity of moments of inertia	238
3.19.13.5	Moments of inertia: examples	238
3.20	Motion of rigid bodies	248
3.20.1	Translational and rotational motion of rigid bodies	248
3.20.2	Rolling motion	251
3.20.3	Precession	257
3.20.3.1	Precession under an applied torque	257
3.20.3.2	Precession of a heavy top	259
3.20.3.3	Precession of the earth's axis of rotation	260
3.20.3.4	Free precession	262
3.21	Rotating frames of reference	263
3.22	Reduction of a system of forces	268
3.22.1	Introduction	268
3.22.1.1	Concurrent forces	268
3.22.1.2	Non-concurrent forces	269
3.22.2	Concurrent systems in equilibrium	270
3.22.2.1	Two concurrent forces in equilibrium	270
3.22.2.2	Three concurrent forces in equilibrium	270
3.22.2.3	More than three concurrent forces	272
3.22.2.4	Moment of concurrent forces in equilibrium	274
3.22.3	Reduction of a system of forces acting on a rigid body	275
3.22.3.1	Reduction of a pair of like parallel forces	275
3.22.3.2	Unlike and unequal parallel forces	277

CONTENTS

3.22.3.3	Equal and unlike parallel forces : couple	277
3.22.3.4	Composition of couples	282
3.22.3.5	A couple and a force in a parallel plane	283
3.22.3.6	Reduction of a system of co-planar forces	284
3.22.3.7	Reduction of non-coplanar forces: wrench	285
3.23	Static and dynamic friction	287
3.23.1	Introduction	287
3.23.2	Static friction	289
3.23.3	Dynamic friction	290
3.23.4	Indeterminate problems in statics: the ladder problem	297
3.23.4.1	The ladder problem	299
3.23.5	The mechanism underlying friction	301
3.23.6	Wet friction and lubrication	310
3.23.7	Rolling friction	311
3.24	Motion of a wheel under driving	316
3.24.1	Driving by means of a couple	316
3.24.2	Driving by means of a force	320
4	Simple Harmonic Motion	324
4.1	Oscillatory motion	324
4.1.1	Simple harmonic motion	325
4.1.1.1	The equation of motion	325
4.1.1.2	Solving the equation of motion	327
4.1.1.3	General and particular solutions	328
4.1.1.4	Relating the solution to initial conditions	329
4.1.1.5	Periodicity of motion	330
4.1.1.6	The phase	331
4.1.1.7	The amplitude	332
4.1.1.8	Graphical representation of the motion	334
4.2	Energy in simple harmonic motion	335
4.2.1	Potential energy	335
4.2.2	Kinetic energy in simple harmonic motion	337

CONTENTS

4.3	Simple harmonic oscillations of physical quantities	339
4.3.1	Angular oscillations	341
4.4	The pendulum and the spring	347
4.4.1	The simple pendulum	347
4.4.2	The spring	350
4.5	Damped simple harmonic motion	354
4.5.1	Damped SHM: equation of motion	354
4.5.2	Underdamped and overdamped motions	357
4.5.2.1	Underdamped SHM	357
4.5.2.2	Overdamped SHM	359
4.5.3	Damped SHM: dissipation of energy	361
4.6	Forced SHM	363
4.6.1	Energy exchange in forced SHM. Resonance.	367
5	Gravitation	373
5.1	Introduction: Newton's law of gravitation	373
5.1.1	Principle of superposition	375
5.2	Gravitational intensity and potential	378
5.2.1	Gravitational intensity	378
5.2.2	Gravitational potential	381
5.2.3	Gravitational potential: summary	383
5.2.4	'Potential at infinity'.	383
5.2.5	Potential due to a point mass	384
5.2.6	Potential due to a number of point masses	386
5.2.7	Describing a gravitational field	388
5.3	Gauss' principle in gravitation	389
5.3.1	Flux of gravitational intensity	389
5.3.2	Gauss' principle	391
5.3.3	Application: a spherically symmetric body	392
5.3.3.1	Intensity and potential at an external point	393
5.3.3.2	A spherical shell	394
5.3.3.3	Gravitational interaction of two spherical bodies	398

CONTENTS

5.4	Earth's gravitational field: acceleration due to gravity	400
5.4.1	Earth's gravitational field	400
5.4.2	Center of gravity	404
5.4.3	The weight of a body	406
5.4.3.1	Weight as a force of reaction: weightlessness	406
5.4.3.2	Weight reduction due to earth's rotation	409
5.4.4	Escape velocity	410
5.5	The motion of planetary bodies	412
5.5.1	Introduction	412
5.5.2	The equation of motion and the nature of trajectories	413
5.5.3	Kepler's laws	415
5.5.4	Circular orbits in an inverse square field	418
5.5.5	Tidal force	420
5.5.6	The orbit of the moon	423
5.5.7	The motion of a projectile	425
5.6	Gravitation: a broader view	426
6	Elasticity	430
6.1	Introduction: External and internal forces in a body	430
6.2	Strain and stress	431
6.3	Quantitative definition of strain: strain parameters	433
6.3.1	Tensile strain	436
6.3.2	Bulk strain	437
6.3.3	Shear strain	438
6.3.4	Mixed strain	441
6.3.5	Principal axes. Principal components of strain.	441
6.4	Stress in a deformable body	442
6.4.1	Tensile stress	445
6.4.2	Shear stress	446
6.4.3	Bulk stress	447
6.4.4	Mixed stress: principal components of stress	448

CONTENTS

6.5	Stress-strain curve	449
6.5.1	A weightless wire with a load	449
6.5.2	The curve: principal features	450
6.5.3	Elastic and plastic deformations	454
6.6	Stress-strain relations: elastic constants	458
6.6.1	Young's modulus	459
6.6.2	Poisson's ratio	461
6.6.3	Modulus of rigidity	465
6.6.4	Bulk modulus	467
6.6.5	Principle of superposition	467
6.6.6	Relations between the elastic constants	469
6.6.7	Elastic properties of fluids	470
6.7	Strain energy	473
7	Mechanics of Fluids	478
7.1	Introduction: the three states of matter	478
7.2	Fluids in equilibrium	480
7.2.1	Internal forces in a fluid: pressure	480
7.2.2	Pressure in an incompressible liquid	483
7.2.3	Thrust of a fluid	486
7.2.4	Atmospheric pressure	490
7.2.5	Buoyancy: Archimedes' Principle	492
7.2.6	Equilibrium of fully or partly submerged body	493
7.2.6.1	Condition of equilibrium	493
7.2.6.2	Stability of equilibrium	496
7.2.7	Pascal's law: transmission of pressure	505
7.3	Fluids in motion: a few introductory concepts	509
7.3.1	Stream lines: steady flow	510
7.3.2	From laminar flow to turbulence	513
7.3.3	Rotational and irrotational flows	513
7.3.4	Equation of continuity	514
7.3.5	Ideal fluid: equation of motion	515

CONTENTS

7.3.6	Energy conservation: Bernoulli's principle	517
7.3.7	Potential flow	520
7.3.8	Lift and drag forces: a brief introduction	521
7.4	The siphon	522
7.5	Viscosity and fluid flow	529
7.5.1	Introduction	529
7.5.2	Newton's formula for viscous force	529
7.5.2.1	Kinematic viscosity	532
7.5.2.2	Variation of viscosity with temperature	532
7.5.3	Viscosity and transport of momentum	533
7.5.4	Viscosity and turbulence	535
7.5.5	Non-Newtonian fluids	536
7.5.6	Poiseuille's flow	538
7.5.7	The origin of the viscous force: a brief outline	542
7.5.8	The boundary layer	544
7.5.8.1	Laminar boundary layer in Poiseuille's flow	545
7.5.8.2	Boundary layer on a flat plate	546
7.5.8.3	Boundary layer near a curved obstacle: boundary layer separation	549
7.5.9	Stability of fluid flow: vorticity and turbulence	557
7.6	Surface energy and surface tension	563
7.6.1	Introduction	563
7.6.2	Surface energy and surface tension: thermodynamic consid- erations	564
7.6.3	The tendency of a liquid surface to shrink	568
7.6.4	Surface tension as lateral force	569
7.6.5	Angle of contact	570
7.6.6	Pressure difference across a curved liquid surface	574
7.6.7	Capillary rise	577
7.6.8	A few phenomena associated with surface tension	584
7.6.8.1	Seepage of water through soil	584

CONTENTS

7.6.8.2	Formation of raindrops and clouds.	584
7.6.8.3	Pouring oil over rough sea	587
7.6.8.4	Walking on water	588
7.6.8.5	Rayleigh-Plateau Instability: beads on cobweb threads	589
7.6.8.6	Velocity of surface waves.	591
7.6.8.7	Surfactants	593
8	Thermal Physics	596
8.1	Thermodynamic systems and their interactions	596
8.1.1	Adiabatic enclosures: Work	597
8.1.2	Diathermic enclosures: Heat	598
8.1.3	Examples of adiabatic and diathermic enclosures	598
8.2	Thermodynamic equilibrium	599
8.3	Thermodynamic processes	600
8.3.1	Adiabatic process	601
8.3.2	States and processes: thermodynamic state diagram	601
8.3.3	Quasi-static processes	602
8.3.4	Thermal equilibrium	603
8.4	The zeroth law of thermodynamics: Temperature.	603
8.4.1	Explaining the zeroth law	603
8.4.2	Temperature as a thermodynamic variable	604
8.4.3	Empirical scales of temperature	605
8.4.4	The SI scale of temperature	608
8.4.5	The direction of heat flow	608
8.4.6	Thermal reservoir: heat source and heat sink	609
8.4.7	Isothermal processes	610
8.5	Adiabatic processes between given states	611
8.6	The significance of adiabatic work	613
8.7	The quantitative definition of heat	616
8.8	The first law of thermodynamics	617
8.8.1	Equivalence of heat and work	619
8.8.2	Work performed by a gas in a quasi-static process	619

CONTENTS

8.8.3	Summary of the first law.	624
8.9	Intensive and extensive variables	628
8.10	The kinetic theory of gases	629
8.10.1	Macroscopic and microscopic descriptions	629
8.10.2	Mole number	631
8.10.3	The ideal gas	632
8.10.4	Pressure of a gas in kinetic theory	633
8.10.5	The kinetic interpretation of temperature.	638
8.10.6	The ideal gas equation of state	642
8.10.6.1	Ideal gas: isothermal and adiabatic processes . . .	644
8.10.7	Random motion of molecules: Molecular collisions.	647
8.10.7.1	Mean free path	649
8.10.8	Brownian motion	650
8.10.9	Mean speed and most probable speed	651
8.10.10	Maxwell's velocity distribution formula	654
8.10.11	Partial pressure	656
8.11	Reversible and irreversible processes	657
8.12	Entropy: thermodynamic definition	661
8.13	The second law of thermodynamics	665
8.14	Statistical physics: Boltzmann's formula	666
8.14.1	The statistical interpretation of entropy	668
8.15	Heat engines	670
8.15.1	Explaining the idea of a heat engine	671
8.15.2	The efficiency of a heat engine	672
8.15.3	An ideal heat engine: the Carnot cycle	675
8.15.4	The absolute scale of temperature	681
8.15.5	Scales of temperature: summary	682
8.15.6	The Rankine cycle	683
8.15.7	The Otto and Diesel cycles	686
8.15.7.1	The Otto cycle	687
8.15.7.2	The Diesel cycle	688

CONTENTS

8.15.8	Refrigeration	690
8.16	Thermal expansion of solids, liquids, and gases	693
8.17	Thermal expansion of solids	695
8.17.1	Coefficients of expansion of a solid	695
8.17.2	Relation between the three coefficients for a solid	696
8.17.3	Instances of thermal expansion of solids	699
8.18	Thermal stress	700
8.19	Thermal expansion of liquids	702
8.19.1	Apparent expansion and real expansion of a liquid	702
8.19.2	The anomalous expansion of water	705
8.20	Thermal expansion of gases	707
8.21	Calorimetry	709
8.21.1	Thermal capacities of bodies	709
8.21.2	Specific heats of substances	711
8.21.3	Specific heats of an ideal gas	712
8.21.4	Adiabatic and isothermal expansion of gases	715
8.21.5	The fundamental principle of calorimetry	718
8.22	Change of state: phase transition	721
8.22.1	Change of state as phase transition	721
8.22.2	Transition temperature. Latent heat	723
8.22.3	Dependence of transition temperature on the pressure	726
8.22.4	Saturated vapor	727
8.22.5	The coexistence curve: Triple point	728
8.22.6	Gas and vapor	731
8.22.7	Saturated air	731
8.22.8	Saturation pressure and superincumbent pressure	732
8.22.9	Evaporation	734
8.22.10	Relative humidity	736
8.22.11	Dew Point	737

CONTENTS

8.23	Transmission of heat	740
8.23.1	Conduction	740
8.23.1.1	Thermal conductivity	742
8.23.1.2	Thermal diffusivity	746
8.23.1.3	Stationary and non-stationary heat flow	746
8.23.2	Convection	749
8.23.2.1	Natural and forced convection	751
8.23.3	Thermal radiation	751
8.23.3.1	Stefan's law of radiation	755
8.23.3.2	Energy exchange in radiative transfer	757
8.23.3.3	Kirchhoff's principle	759
8.23.3.4	Newton's law of cooling	761
8.23.3.5	The greenhouse effect	763
8.24	Supplement: Random variables and probability distributions	766
9	Wave motion I: Acoustic waves	770
9.1	Simple harmonic oscillations of physical quantities	770
9.2	Oscillations transmitted through space: waves	771
9.3	Sound waves as variations in pressure: elastic waves	773
9.4	Sound waves in one dimension	776
9.4.1	Variation of excess pressure	776
9.4.2	Propagation of the monochromatic wave	779
9.5	Waves in three dimensions	783
9.5.1	The plane progressive wave	783
9.5.2	Waves of more general types	785
9.5.3	The principle of superposition	786
9.6	The wave equation	787
9.7	Sources and boundary conditions	789
9.7.1	The monopole source. Spherical wave.	790
9.7.2	Dipole and quadrupole sources	792
9.7.3	Sources and wave patterns: summary	795
9.8	Reflection and refraction of plane waves	797

CONTENTS

9.9	Diffraction and scattering by obstacles	800
9.9.1	Finite extent of interface	800
9.9.2	Curvature of the interface	802
9.9.3	Wave incident on an uneven surface	803
9.10	Echo and reverberation of sound	805
9.10.1	Echo	805
9.10.2	Reverberation.	807
9.11	Velocity, energy density, and intensity	809
9.11.1	Formulation of the problem	809
9.11.2	Displacement and strain	812
9.11.3	Velocity of sound in a fluid	814
9.11.3.1	Velocity of sound in an ideal gas	816
9.11.3.2	Dependence on pressure, temperature, and humidity	820
9.11.4	Energy density and intensity	822
9.11.4.1	Spherical waves: the inverse square law of intensity.	825
9.12	Ultrasonic waves	828
9.13	Döppler effect	829
9.13.1	Introduction	829
9.13.2	Frequency related to rate of change of phase	832
9.13.3	Döppler effect:the general formula	834
9.13.4	Uniform motions of source and observer	835
9.14	Supersonic objects and shock waves	841
9.14.1	The production of shock fronts	844
9.15	Superposition effects	848
9.15.1	Interference	849
9.15.1.1	Introduction to the idea of coherence	849
9.15.1.2	Interference as superposition of coherent waves	851
9.15.2	Standing waves	854
9.15.2.1	Standing waves in an air column	854
9.15.2.2	Features of a standing wave	855

CONTENTS

9.15.2.3	Superposition of propagating waves	857
9.15.2.4	Progressive waves and standing waves: a few points of distinction	858
9.15.2.5	Modes of standing waves in an air column	859
9.15.3	Beats	862
9.15.4	Wave packets: group velocity	864
9.16	Vibrations of strings and diaphragms	867
9.16.1	Transverse vibrations of stretched strings	868
9.16.2	Vibrations of stretched diaphragms	874
9.16.3	Musical instruments	878
9.17	Loudness, pitch, and quality of sound	879
9.18	Elastic waves in solids	880
9.18.1	Vibrations in a crystalline medium: normal modes	881
9.18.2	Elastic waves in an isotropic solid	882
10	Ray Optics	886
10.1	Introduction	886
10.2	Ray optics: basic principles	887
10.3	Image formation by rays originating from a point source	898
10.4	Image formation by reflection at a plane surface	901
10.5	Refraction at a plane surface	905
10.5.1	Image formation	905
10.5.2	Refraction through a layer with parallel surfaces	911
10.6	Total internal reflection	913
10.7	Prism	918
10.7.1	The basic formulae	919
10.7.2	Deviation	920
10.7.3	Limiting angle of incidence	922
10.7.4	Minimum deviation	923
10.8	Reflection and refraction at spherical surfaces	925
10.8.1	Spherical mirrors: a few definitions	925
10.8.2	Refraction at a spherical surface: definitions	929

CONTENTS

10.9	Sign convention in ray optics	932
10.10	Image formation in reflection by a spherical mirror	936
10.10.1	Focal length of a spherical mirror	936
10.10.2	Aperture	938
10.10.3	Image formation: relation between object distance and image distance	939
10.10.3.1	Image for an off-axis object point	943
10.10.3.2	Image formation for short extended objects	945
10.11	Image formation by refraction at a spherical surface	946
10.11.1	Refraction at a spherical surface: image formation for a point object on the axis	948
10.11.2	Image formation for a short extended object	953
10.12	Spherical lens	955
10.12.1	Image formation by a thin lens	961
10.12.2	Real and virtual image formation by a convex lens	967
10.12.3	Image formation for off-axis points	968
10.12.4	Longitudinal and transverse magnifications	972
10.12.5	Angular magnification	973
10.12.6	Minimum distance between object and real image	975
10.13	Combination of thin lenses	977
10.13.1	Equivalent lens	979
10.13.2	Thin lenses in contact	982
10.14	Aberrations in image formation	983
10.14.1	Monochromatic and chromatic aberrations	984
10.14.2	Types of monochromatic aberration	986
10.14.3	Aberrations: an overview	986
10.14.4	Correcting an optical system for aberrations	988
10.14.5	Image imperfection: aberration and diffraction	988
10.15	The human eye	989
10.16	Optical instruments	991
10.16.1	The camera	991

CONTENTS

10.16.2	The telescope and the compound microscope	994
10.16.2.1	The telescope	994
10.16.2.2	The compound microscope	996
11	Electrostatics	1000
11.1	Introduction: elementary charges	1000
11.2	Acquisition of charges by bodies	1002
11.2.1	Transfer of elementary charges	1002
11.2.2	Contact electrification and contact potential	1004
11.3	Electrostatic force between charges	1006
11.3.1	Coulomb's law	1006
11.3.2	The principle of superposition	1009
11.4	Electric field intensity and potential	1011
11.4.1	Electric field intensity	1011
11.4.2	Electrical potential	1014
11.4.3	Electrical potential: summary	1016
11.4.4	Potential 'at infinity'	1016
11.4.5	Potential due to a point charge	1019
11.4.6	Potential due to a number of point charges	1021
11.4.7	Deriving the intensity from the potential	1023
11.5	Force on a thin layer of charge	1026
11.6	Electric dipole and dipole moment	1028
11.6.1	A pair of equal and opposite point charges	1028
11.6.2	Dipole moment and dipole	1030
11.6.3	Torque and force on a dipole in an electric field	1036
11.6.3.1	Torque on a dipole	1036
11.6.3.2	Force on a dipole	1038
11.6.4	Potential energy of a dipole in an electric field	1039
11.7	Electric lines of force and equipotential surfaces	1042
11.7.1	Geometrical description of an electric field. Neutral points. . .	1042
11.7.2	Characteristics of lines of force	1045
11.7.3	Equipotential surfaces	1047

CONTENTS

11.7.4	Density of lines of force. Tubes of force.	1049
11.7.5	Separations between equipotential surfaces	1051
11.8	Gauss' principle in electrostatics	1052
11.8.1	Flux of electric field intensity	1052
11.8.2	Gauss' principle	1054
11.8.2.1	Flux due to a single point charge	1055
11.8.2.2	Solid angle	1055
11.8.2.3	Gauss' principle: derivation	1058
11.9	Applications of Gauss' principle	1059
11.9.1	A charged spherical conductor	1059
11.9.2	A spherically symmetric charge distribution	1061
11.9.3	Potential energy of a uniformly charged sphere	1063
11.9.4	An infinitely long cylindrical conductor	1064
11.9.5	An infinitely extended planar sheet of charge	1066
11.10	Conductors and dielectrics	1068
11.10.1	Free and bound electrons	1068
11.10.2	Electric field intensity and charge density within a conductor .	1069
11.10.3	Conductor: surface charge density.	1070
11.10.3.1	Field intensity on the surface of a charged conductor	1073
11.10.3.2	Force on the surface of a charged conductor	1075
11.10.4	Accumulation of charge at sharp points	1076
11.10.5	Polarization in a dielectric medium	1079
11.10.5.1	The polarization vector. Electric susceptibility. . . .	1080
11.10.5.2	Electric field intensity and the displacement vectors	1081
11.10.5.3	Field variables: the question of nomenclature	1083
11.10.5.4	Electric field in a dielectric: summary	1084
11.10.5.5	Field variables as space- and time averages	1085
11.10.5.6	A brief note on relative permittivity	1085
11.11	Capacitors and capacitance	1086
11.11.1	Charges and potentials on a pair of conductors	1086
11.11.2	The Uniqueness Theorem	1090

CONTENTS

11.11.3	Capacitance of a pair of conductors	1092
11.11.4	The spherical condenser	1094
11.11.5	A pair of concentric spherical conductors	1095
11.11.6	Capacitance of a single conductor	1098
11.11.7	Self-capacitance and mutual capacitance	1099
11.11.8	The parallel plate capacitor	1100
11.11.8.1	Charge distribution in the plates	1102
11.11.9	Cylindrical capacitor	1106
11.11.10	Energy of a system of charged conductors	1107
11.11.10.1	Energy of a single charged conductor	1107
11.11.10.2	Energy of a charged parallel plate capacitor	1110
11.11.10.3	Electric field energy	1111
11.11.11	Capacitors in series and parallel	1111
11.11.12	Capacitors with dielectrics	1116
11.12	The potential of the earth	1119
12	Electricity I: Steady currents and their magnetic effects	1121
12.1	Electrical Cells	1121
12.1.1	The half cell	1121
12.1.2	Electrochemical cell	1123
12.1.2.1	Half-cell potential	1123
12.1.2.2	Electromotive force of an electrochemical cell . . .	1124
12.1.2.3	The Galvanic cell	1125
12.1.2.4	The electrolytic cell	1127
12.1.2.5	Primary and secondary cells	1129
12.2	Electrical conductors and electric current	1131
12.3	The current set up by a Galvanic cell	1138
12.4	Ohm's law. Electrical units	1143
12.4.1	Current density and current	1143
12.4.2	Resistance and resistivity	1144
12.4.3	Temperature dependence of resistivity	1150

CONTENTS

12.5	Steady current in a conductor produced by an electrical cell	1151
12.5.1	Transformation of energy	1151
12.5.2	The pathway of energy flow	1156
12.5.3	Electromotive force (EMF) and source of EMF	1159
12.5.4	Heating effect of current: Joule's law of heating	1163
12.5.5	Summary: electrical cells, EMFs, and currents	1165
12.6	Series and parallel combination of resistances	1168
12.6.1	The laws of series and parallel combination	1168
12.6.2	Voltage division and current division	1174
12.6.2.1	Voltage division	1174
12.6.2.2	Current division	1176
12.7	Analysis of DC electrical circuits	1177
12.7.1	Kirchhoff's principles	1178
12.7.1.1	Kirchhoff's first principle	1178
12.7.1.2	Kirchhoff's second principle	1179
12.7.2	The principle of superposition	1181
12.7.3	The Wheatstone bridge	1183
12.8	The magnetic effect of currents	1191
12.8.1	Force between currents composed from elementary forces . . .	1192
12.8.2	Force between a pair of parallel current-carrying wires	1196
12.8.3	Magnetic field intensity	1197
12.8.4	The force on a moving charge in a magnetic field	1200
12.8.5	Field variables: the question of nomenclature	1202
12.8.6	Field due to a current loop. Principle of superposition.	1202
12.8.7	Magnetic lines of force	1205
12.8.8	Field intensity due to a straight wire	1207
12.8.8.1	Infinitely long and straight wire	1208
12.8.9	Magnetic field intensity due to a circular wire	1211
12.8.9.1	Magnetic field of a solenoid	1214
12.8.10	Ampere's circuital law	1218

CONTENTS

12.8.11	Applications of the circuital law	1220
12.8.11.1	Ampere's law: field due to a long straight wire . . .	1220
12.8.11.2	Ampere's law: the tightly wound long solenoid . . .	1222
12.8.11.3	Infinitely long cylindrical current distribution . . .	1224
12.8.12	The magnetic dipole	1225
12.8.12.1	Electric and magnetic dipole moments	1227
12.8.12.2	Current loop: a surface distribution of dipoles . . .	1229
12.8.12.3	Torque and force on a magnetic dipole	1232
12.8.12.4	Energy of a magnetic dipole in a magnetic field . .	1233
12.8.13	Magnetic field: comparison with electrostatics	1234
12.8.14	Currents and magnetic fields: overview	1236
12.9	Magnetic properties of materials	1239
12.9.1	Magnetization in a material body	1240
12.9.2	Magnetic susceptibility and magnetic permeability	1241
12.9.2.1	Field variables as space- and time averages	1244
12.9.3	Dia- and paramagnetism	1245
12.9.3.1	Paramagnetism	1246
12.9.3.2	Diamagnetism	1250
12.9.4	Ferromagnetism	1255
12.9.4.1	Spontaneous magnetization	1255
12.9.4.2	Magnetic domains	1257
12.9.4.3	The magnetization curve: hysteresis	1259
12.9.4.4	Residual magnetism: permanent magnets	1260
12.9.4.5	Transition to paramagnetic behaviour	1263
12.10	The earth as a magnet: geomagnetism	1264
12.11	The chemical effect of current	1267
12.11.1	Electrolytes and electrolysis	1267
12.11.2	Faraday's laws of electrolysis	1269
12.12	Thermoelectric effects	1270
13	Electricity II: Varying and alternating currents	1273
13.1	Introduction	1273

CONTENTS

13.2	Electromagnetic induction	1275
13.2.1	Magnetic flux	1277
13.2.2	Faraday's law of electromagnetic induction	1279
13.2.2.1	Lenz's law	1281
13.2.2.2	Motional EMF	1282
13.2.3	The principle of DC and AC generators	1286
13.2.3.1	Conducting frame rotating in a magnetic field . . .	1287
13.2.4	Rotating magnetic field: AC motors	1292
13.2.4.1	The synchronous motor	1294
13.2.4.2	The asynchronous motor	1295
13.2.5	The principle of DC motors. Back EMF.	1296
13.2.5.1	DC motor: back EMF	1299
13.2.6	Self-inductance	1302
13.2.6.1	Self-inductance of a long solenoid	1303
13.2.6.2	Self-inductance of a toroidal solenoid	1304
13.2.6.3	Inductor	1305
13.2.6.4	Back EMF in an inductor	1306
13.2.7	Mutual inductance	1306
13.3	Varying currents in electrical circuits	1308
13.3.1	Currents and voltages in an L - R circuit	1310
13.3.1.1	Growth of current	1310
13.3.1.2	Decay of current	1313
13.3.2	Analysis of circuits with varying currents	1314
13.3.3	Currents and voltages in a C - R circuit	1317
13.3.3.1	Growth of charge	1317
13.3.3.2	Decay of charge	1319
13.3.4	Oscillations in an L - C - R circuit	1320
13.4	Magnetic field energy	1322
13.5	Alternating currents	1326
13.5.1	Mathematical description of AC currents and voltages	1326
13.5.1.1	Amplitude, frequency, and phase	1327

CONTENTS

13.5.1.2	Root mean squared values	1327
13.5.1.3	The complex representation of AC quantities	1329
13.5.2	An L - C - R circuit with an AC source	1331
13.5.3	Impedance	1336
13.5.4	Analysis of AC circuits	1339
13.5.5	Power in an AC circuit	1344
13.5.6	The three-phase supply	1347
13.5.7	The transformer	1353
13.5.7.1	Back EMFs	1354
13.5.7.2	The loading of the primary	1356
13.5.7.3	The current ratio	1357
13.5.7.4	Energy losses in the transformer	1358
13.5.7.5	The transformer in three phase distribution	1359
13.5.8	Eddy currents	1360
14	Wave motion II: Electromagnetic waves	1365
14.1	Introduction	1365
14.2	Electromagnetic theory	1366
14.2.1	The electromagnetic field in free space	1366
14.2.1.1	What the first equation means	1366
14.2.1.2	What the second equation means	1368
14.2.1.3	What the third equation means	1368
14.2.1.4	What the fourth equation means	1369
14.2.1.5	The four equations: an overview	1371
14.2.2	Electromagnetic fields in material media	1372
14.3	Electromagnetic waves	1373
14.3.1	Sources of electromagnetic waves	1374
14.3.2	Transmission of energy	1375
14.3.3	The principle of superposition	1375
14.4	The plane progressive monochromatic wave	1377
14.4.1	Space-time field variations	1377
14.4.2	Frequency, wavelength and velocity	1380

CONTENTS

14.4.3	The phase	1382
14.4.4	The wave front and its propagation	1384
14.4.5	The electromagnetic spectrum	1385
14.4.6	Energy flux and intensity	1386
14.4.6.1	Energy density and energy flux	1386
14.4.6.2	Intensity	1389
14.4.6.3	Velocity of energy transport	1391
14.4.7	Radiation pressure	1394
14.4.8	The state of polarization of an electromagnetic wave	1396
14.4.9	Wave propagating in an arbitrarily chosen direction	1399
14.4.10	Wave normals and rays	1401
14.5	The complex representation of wave functions	1402
14.6	Reflection and refraction of plane waves	1404
14.6.1	Reflection	1407
14.6.2	Refraction	1408
14.6.3	Total internal reflection	1409
14.7	Dispersion and absorption	1410
14.7.1	Plane waves in a dielectric medium	1410
14.7.1.1	Features of dielectric constant: summary	1414
14.7.2	Plane waves in a conducting medium: attenuation	1415
14.7.3	Negative refractive index: metamaterials	1418
14.8	The monochromatic spherical and cylindrical waves	1419
14.9	Wave packet and group velocity	1422
14.10	Coherent and incoherent waves	1423
14.11	Stationary waves	1426
15	Wave Optics	1430
15.1	Introduction	1430
15.2	Experiments with an illuminated aperture	1432
15.2.1	Spreading and bending of waves	1435
15.2.2	Waves coming out of pin-holes and slits	1437

CONTENTS

15.3	Interference of coherent waves	1439
15.3.1	Superposition of two plane waves	1439
15.3.1.1	The resultant intensity.	1442
15.3.1.2	Maxima and minima in $I(\mathbf{r})$	1442
15.3.1.3	Conditions for interference	1444
15.3.2	A simplified approach: interference of scalar waves	1445
15.3.3	The complex representation of wave functions	1446
15.3.4	Young's pattern with a pair of pin-holes	1448
15.3.4.1	Phase difference and path difference	1451
15.3.5	Young's pattern with a pair of slits	1454
15.3.6	Young's fringes with partially coherent light	1457
15.3.6.1	Young's pattern with unpolarized light	1457
15.3.6.2	Quasi-monochromatic light	1458
15.3.6.3	Coherence time	1459
15.3.6.4	Coherence length	1460
15.3.7	Thin film patterns	1461
15.3.7.1	Fringes of equal inclination	1467
15.3.7.2	Fringes of equal thickness	1469
15.3.7.3	The color of thin films	1474
15.3.7.4	Non-reflective coatings	1478
15.4	Diffraction of light	1479
15.4.1	Introduction	1479
15.4.2	The basic approach in diffraction theory	1481
15.4.3	The intensity distribution	1484
15.4.4	Fraunhofer and Fresnel diffraction patterns	1487
15.4.5	The single slit Fraunhofer pattern	1489
15.4.5.1	Ray paths corresponding to secondary waves.	1489
15.4.5.2	The intensity formula.	1491
15.4.5.3	Absence of diffraction in the vertical direction.	1492
15.4.5.4	The intensity graph	1493
15.4.5.5	Fraunhofer fringes with a slit-source.	1496

CONTENTS

15.4.5.6	Phase in Fraunhofer diffraction	1497
15.4.5.7	Coherence properties and diffraction fringes	1498
15.4.6	The double slit Fraunhofer pattern	1499
15.4.7	The diffraction grating	1504
15.4.8	Resolving powers of optical instruments	1507
15.5	Polarized and unpolarized light	1510
15.5.1	The basic components: x-polarized and y-polarized light	1510
15.5.2	Specifying the basic components and their phase relation	1511
15.5.3	Correlations: polarized and unpolarized light	1512
15.5.4	Elliptically polarized light	1512
15.5.5	Circularly polarized and linearly polarized light	1513
15.5.6	Intensity relations	1515
15.5.7	Optical anisotropy: double refraction	1516
15.5.8	Production of polarized light	1517
15.6	Lasers: coherent sources of light	1518
15.6.1	Emission and absorption as quantum processes	1519
15.6.2	The state of a photon	1520
15.6.3	Classical and quantum descriptions of the field	1521
15.6.4	Stimulated emission of radiation	1522
15.6.5	Stimulated emission and coherent waves	1523
15.6.6	Population inversion	1525
15.6.7	Light amplification in a resonant cavity	1527
15.6.8	The laser as a coherent source of light: summary	1528
15.7	Holography	1529
15.8	Scattering of light	1535
15.8.1	Rayleigh scattering	1535
15.8.1.1	Rayleigh scattering by a single scatterer	1536
15.8.1.2	Rayleigh scattering in a fluid	1540
15.8.2	Mie scattering	1542
15.8.3	Raman scattering	1543
15.9	Wave optics and ray optics	1544

CONTENTS

16 Quantum theory	1549
16.1 Introduction	1549
16.1.1 Quantum and classical concepts: analogy from optics	1550
16.1.2 Emergence of quantum concepts	1551
16.2 Quantum and classical descriptions of the state of a system	1552
16.2.1 Illustration: the free particle in one dimension	1554
16.2.2 Wave-like features	1556
16.2.3 Wave function: de Broglie relations	1557
16.2.4 Quantum description of state: summary	1557
16.3 The principle of uncertainty	1559
16.3.1 Uncertainty in momentum	1559
16.3.2 Momentum and position uncertainties	1561
16.4 Observable quantities, probability distributions, and uncertainties	1562
16.5 The simple harmonic oscillator	1568
16.5.1 Bound system: quantization of energy	1569
16.5.2 Digression: the continuous and the discrete	1570
16.5.3 Harmonic oscillator: the uncertainty principle at work	1571
16.6 Time evolution of states	1574
16.7 Superposed states in quantum theory	1576
16.8 Mixed states: incoherent superposition	1577
16.9 Black body radiation: Planck's hypothesis	1578
16.9.1 Harmonic oscillators in thermal equilibrium	1582
16.10 Bohr's theory of the hydrogen atom	1584
16.10.1 The hydrogen spectrum	1585
16.10.2 Bohr's postulates and the hydrogen spectrum	1587
16.10.3 Bohr's theory and the quantum theory of the atom	1590
16.10.4 The hydrogen spectrum: mechanism	1592
16.11 Applications of Bohr's theory	1594
16.12 Bound and unbound systems: standing and traveling waves	1595
16.13 Photoelectric effect: Einstein's theory	1596
16.13.1 Features of photoelectric emission	1597

CONTENTS

16.13.2	The role of photons in photoelectric emission	1599
16.13.3	Bound systems and binding energy	1600
16.13.4	The basic equation for photoelectric emission	1603
16.14	The Compton effect	1607
16.15	Quantum theory goes deep: particles and fields	1611
17	Relativity: the special and the general theory	1614
17.1	Relativity: Introduction	1614
17.1.1	Introduction: frames of reference, inertial frames	1614
17.1.2	Introduction: the Galilean principle of equivalence	1615
17.1.3	Introduction: the non-relativistic and the relativistic	1616
17.1.4	Introduction: the equivalence principles	1618
17.2	The special theory of relativity	1619
17.2.1	Inertial frames and the velocity of light	1619
17.2.2	The Lorentz transformation formulae	1620
17.2.3	Space-time interval	1624
17.2.4	Lorentz transformation: the general form	1627
17.2.5	Consequences of the Lorentz transformation formula	1631
17.2.5.1	Relativity of simultaneity	1633
17.2.5.2	Lorentz contraction	1634
17.2.5.3	Time dilatation	1637
17.2.5.4	Velocity transformation	1640
17.2.5.5	Relativistic aberration	1643
17.2.6	Space-time diagrams and world lines	1644
17.2.6.1	Representation of events and world lines	1644
17.2.6.2	The space-time diagram and Lorentz transformations	1646
17.2.6.3	The invariant regions	1648
17.2.6.4	Time-like and space-like separations	1650
17.2.7	Space-time geometry	1651
17.2.7.1	Geometry in '1+1' dimensions	1651
17.2.7.2	The (1+3)-dimensional space-time geometry	1654

CONTENTS

17.2.8	Physical quantities as four-vectors	1656
17.2.8.1	Vectors: the basic idea	1656
17.2.8.2	Four-vectors	1659
17.2.8.3	Four-vectors and tensors: a primer	1660
17.2.8.4	The velocity four-vector	1667
17.2.8.5	Relativistic mass, relativistic momentum, and relativistic energy	1669
17.2.8.6	The energy-momentum four-vector	1672
17.2.8.7	The Doppler effect	1674
17.2.8.8	The force four-vector	1678
17.2.9	The electromagnetic field as a tensor	1681
17.3	The general theory of relativity: a brief introduction	1685
17.3.1	Introduction: the general principle of equivalence	1685
17.3.2	Tensor fields	1688
17.3.3	Einstein's equation for the metric tensor	1690
17.3.4	Equation of motion in a gravitational field	1692
17.3.5	Gravitation and the electromagnetic field	1693
17.3.6	The Schwarzschild solution	1694
17.3.7	Schwarzschild solution: a few consequences	1697
17.3.7.1	The Newtonian limit	1697
17.3.7.2	Gravitational time dilatation and red shift	1698
17.3.7.3	Black holes	1701
17.3.8	The general theory of relativity: the classical and the quantum	1702
18	Atoms, Nuclei, and Molecules	1704
18.1	Introduction	1704
18.2	The atomic nucleus: atomic volume and mass	1705
18.3	Single-electron states	1707
18.3.1	Single-electron states: elliptic orbits and degeneracy	1709
18.3.2	Single-electron states: space quantization	1712
18.3.3	Single-electron states: electron spin	1713
18.3.4	Single-electron states: summary and notation	1715

CONTENTS

18.4	Building up the atom	1716
18.4.1	Electronic configuration and electron shells	1716
18.4.2	Electronic configurations and the periodic table	1720
18.5	The atom as a whole	1721
18.5.1	Screening of the nuclear charge	1721
18.5.2	Quantum theory of atomic states: a brief outline	1724
18.5.2.1	The indistinguishability principle and its consequences	1724
18.5.2.2	Electron-electron interaction: the central field . . .	1726
18.5.2.3	Electron-electron interaction: the spin-dependent residual term	1728
18.5.2.4	Spin-orbit coupling: excited states of sodium and magnesium	1729
18.5.3	The atom as a whole: summary and overview	1734
18.6	Continuous and characteristic X-ray spectra	1735
18.6.1	Bohr's theory and X-ray spectra	1737
18.7	Atomic spectra	1741
18.8	Physics of the atomic nucleus	1743
18.8.1	The atomic number and the mass number	1743
18.8.2	The nucleon: internal characteristics	1743
18.8.3	The interaction force between nucleons	1745
18.8.3.1	The saturation property of nuclear forces	1747
18.8.3.2	The nucleus as a liquid drop: nuclear radius	1749
18.8.4	Nuclear binding energy and mass: nuclear stability	1751
18.8.4.1	The mass-energy equivalence principle	1751
18.8.4.2	Units for nuclear masses	1752
18.8.4.3	Relating nuclear mass to binding energy	1753
18.8.4.4	Binding energy and nuclear stability	1754
18.8.5	The binding energy curve	1755
18.8.5.1	The starting point	1755
18.8.5.2	Finite size effect: the surface correction	1757

CONTENTS

18.8.5.3	The effect of the nuclear charge	1758
18.8.5.4	Other corrections: the mass formula	1758
18.8.5.5	The graph	1760
18.8.6	Single particle and collective nuclear excitations	1762
18.8.7	Radioactive decay	1765
18.8.7.1	Alpha decay	1766
18.8.7.2	Beta decay	1768
18.8.7.3	Gamma decay	1772
18.8.7.4	Radioactive decay law	1773
18.8.7.5	Successive radioactive disintegrations	1775
18.8.8	Nuclear reactions	1776
18.8.8.1	Introduction: examples of nuclear reactions	1776
18.8.8.2	Conservation principles in nuclear reactions	1778
18.8.8.3	Energy balance in nuclear reactions	1778
18.8.8.4	Nuclear fission	1780
18.8.8.5	Nuclear fusion	1785
18.8.9	Introduction to elementary particles	1787
18.8.9.1	The classification of elementary particles	1788
18.8.9.2	Elementary particles and quantum numbers	1790
18.8.9.3	Anti-particles	1792
18.8.9.4	The quark structure of elementary particles	1793
18.8.9.5	The basic interactions	1794
18.8.9.6	The conservation principles	1795
18.8.9.7	The mediating particles	1796
18.8.9.8	Symmetries and the conservation laws	1798
18.8.9.9	The Higgs field and the Higgs boson	1798
18.9	The physics of molecules	1799
18.9.1	The binding of atoms in molecules: molecular bonds	1800
18.9.1.1	The ionic bond	1800
18.9.1.2	The covalent bond	1802
18.9.1.3	The hydrogen bond	1805

CONTENTS

18.9.2	Stationary states of molecules: molecular excitations	1806
18.10	From molecules to solids	1811
19	Electronics	1813
19.1	Introduction	1813
19.2	Electrical properties of semiconductors	1815
19.2.1	Energy bands of electrons in a crystal	1815
19.2.2	The filling up of the bands	1818
19.2.3	The valence- and the conduction bands	1819
19.2.4	Electrochemical potential and the Fermi level	1823
19.2.5	Energy bands: summary	1825
19.2.6	Conductors, insulators, and intrinsic semiconductors	1826
19.2.7	Doped semiconductors	1828
19.2.8	Intrinsic and doped semiconductors: summary	1833
19.3	The p-n junction diode	1833
19.3.1	The junction diode: structural features	1834
19.3.2	The junction diode at thermal equilibrium	1835
19.3.3	The junction diode in forward and reverse bias	1839
19.3.4	Junction diode: current-voltage graph	1841
19.3.5	Junction diode: summary	1843
19.3.6	The diode as rectifier	1844
19.3.7	Special-purpose diodes	1848
19.3.7.1	The Zener diode	1848
19.3.7.2	The light emitting diode	1851
19.3.7.3	The laser diode	1853
19.4	The bipolar junction transistor	1855
19.4.1	The emitter, the base, and the collector	1855
19.4.2	The two-diode model of the transistor	1857
19.4.3	Transistor currents and voltages	1857
19.4.4	The transistor in the active mode	1858
19.4.5	The transistor in the saturation and cut-off modes	1859
19.4.6	Transistor characteristics	1860

CONTENTS

19.4.7	The parameters α and β of the transistor	1864
19.4.8	Convention for using notations	1866
19.4.9	The common emitter input impedance	1866
19.4.10	AC transistor operation: summary	1867
19.4.11	Voltage amplification	1868
19.4.11.1	DC bias: the Q-point	1868
19.4.11.2	Blocking and bypass capacitors	1870
19.4.11.3	AC operation of the amplifier: voltage gain	1871
19.5	The operational amplifier (Op-amp)	1874
19.5.1	The differential amplifier	1875
19.5.2	Op-amp basics	1876
19.6	Oscillators	1879
19.7	Introduction to digital electronics	1880
19.7.1	Boolean algebra	1881
19.7.1.1	Digital circuits and binary numbers	1882
19.7.1.2	Combinational and sequential circuits	1883
19.7.2	The basic logic gates	1884
19.7.2.1	The OR and AND gates with diodes	1886
19.7.2.2	The NOT gate with a transistor	1887
19.7.3	Logic families	1889
19.7.4	The Exclusive-OR, NOR, and NAND gates	1889
19.7.5	Boolean identities and Boolean expressions	1891
19.7.5.1	De Morgan's identities	1891
19.7.6	The binary numbers. Binary arithmetic	1892
19.7.6.1	The Decimal, Binary, Octal, and Hex systems	1892
19.7.6.2	Bits and Bytes	1893
19.7.7	Eight-bit arithmetic	1894
19.7.7.1	1's complement and 2's complement	1894
19.7.7.2	Addition and subtraction in 8-bit arithmetic	1895
19.7.7.3	Overflow and carry	1897
19.7.7.4	Binary multiplication and division	1897

CONTENTS

19.7.8	The adder	1898
19.7.9	Flip-flops	1900
19.7.9.1	the SR flip-flop	1900
19.7.9.2	the D flip-flop	1903
19.7.9.3	The JK flip-flop	1905

Chapter 1

Introduction: Units and Dimensions

1.1 Physical quantities and their units

The physical sciences seek to explain natural phenomena and phenomena observed in the laboratory in *quantitative* terms. For this, one refers to appropriately defined *physical quantities* that describe desired properties or characteristics of systems under study. For instance, the instantaneous velocity of a particle, or the concentration of a compound involved in a chemical reaction, are attributes that can be measured with appropriate equipment, and expressed quantitatively, where the quantitative expression carries some definite information about the system under consideration. Concepts like love and anger, on the other hand, cannot be given quantitative and measurable expression and cannot be included in the category of physical quantities, though these too can be expressed in a way so as to carry some information.

The quantitative expression of a physical quantity is usually made up of two parts - a numerical *value*, and a *unit*. The unit distinguishes between physical quantities of various different *types*. For instance, the mass and the energy of a particle are physical attributes of distinct from each other, expressing two different aspects of its existential and dynamical behaviour, corresponding to which the two are described in terms of

distinct units.

Suppose that one measures the density of water in a laboratory with appropriate equipment and comes up with the result $1000 \text{ kg}\cdot\text{m}^{-3}$. This result tells us that the unit of density being used is $1 \text{ kg}\cdot\text{m}^{-3}$, while the density of water is 1000 times this unit. On further analyzing the unit in this instance, it is found to be made up of two simpler units, namely, a kilogram (kg) and a meter (m).

The unit used in expressing a physical quantity is nothing but a conventionally accepted standard quantity of the same category, while its numerical value tells us in quantitative terms how the quantity under consideration compares in magnitude with that standard quantity. For instance it may be accepted by common agreement that a particular body kept under specified conditions will act as the standard of mass, being called, say, 1 kg. Supposing that, in addition, a procedure has been formulated that enables one to measure how the mass of any other body compares quantitatively with that of the body chosen as the standard, one can use the kg as the unit of mass. Evidently, instead of the body chosen by common agreement as the standard of mass, some *other* body could be chosen and a different unit of mass could be established. In other words, the unit of a physical quantity is not unique, and there can be numerous alternative units for any physical quantity of a particular type. For instance, the gram, the pound, and the kilogram, are three possible units of mass, while other units are also possible.

1.2 Systems of units: the SI system

1.2.1 Relations among physical quantities, and their units

In the physical sciences, one chooses a *system* of units from among numerous possible systems, a commonly used system in recent times being the *standard international* (SI) system. A system of units involves a *consistent* assignment of units to physical quantities of various different types.

While some physical quantities are defined by direct reference to empirical observations,

some others are defined in terms of other quantities of different types. Apart from these defining relations, numerous other relations can be derived among physical quantities of various descriptions by theoretical analysis combined with experimental observations. In other words, there exists a vast network of *relations among physical quantities*.

A consistent assignment of units has to reflect properly the relations, arrived at either by empirical observations or by theoretical reasoning (or, more commonly, by both), among all the physical quantities in use. For instance, in the SI, system, the units of distance and time being, respectively, the meter (m) and the second (s), the unit of velocity, which is defined as a ratio of distance and time, has to be meter per second ($\text{m}\cdot\text{s}^{-1}$), and not, say, meter per kilogram. It is to be mentioned that the relations between various physical quantities do not depend on the system of units chosen and, on the contrary, the assignment of units in any and every system has to be consistent with these relations. For instance, in the *cgs* system, the units of distance and time being, respectively, the centimeter (cm) and the second, the unit of velocity is $\text{cm}\cdot\text{s}^{-1}$, corresponding once again to the fact that the velocity is defined as a ratio of distance and time.

1.2.2 The dimension of a physical quantity

The fact that, whatever the system of units chosen, the unit of any given physical quantity has to reflect properly the nature of that quantity (as revealed by its relations with other physical quantities), is expressed by saying that the unit must be consistent with the *dimension* of that quantity. The dimension expresses the type or category of the physical quantity under consideration from which there follows its unit in any chosen system. Put differently, the dimension expresses the way the unit of the physical quantity under consideration relates to units of other quantities.

However, the dimension does not completely specify the type of a quantity since a quantity is completely specified only through its conceptual interrelation with other measurable quantities. Thus, it may so happen that *different* physical quantities, i.e., ones defined in terms of distinct concepts, *correspond to the same dimension*. For instance,

work and *torque* are two distinct physical quantities, having the *same* dimension which, moreover, is also the same for *energy*. While work and energy are related concepts, they correspond nevertheless to quite distinct conceptual categories.

Before I can explain all this with the help of examples of dimensions and units of some of the physical quantities that we will come across in this book, I have to tell you something about *basic* and *derived* units.

1.3 Basic and derived units

As in the example of the unit of velocity which is made up of units of distance and time, one finds that the unit of all physical quantities can be expressed in terms of a relatively small number of *basic* units. These basic units can be chosen in more ways than one. However, the units of *distance*, *mass*, and *time* are commonly chosen as the basic units, in terms of which the units of numerous other physical quantities are expressed. In addition, the SI system makes use of a number of other basic units (in the context of the SI system and its accepted terminology, these are referred to as *base* units) so as to conveniently express the units of all the other physical quantities in terms of a total of *seven* basic (or base) units.

Units of other physical quantities that can be expressed in terms of these base units are referred to as *derived* units. The SI system includes a list of twenty two special derived units for which distinctive names have been given. All other units are then expressed conveniently by making use of these twenty two derived units along with the seven base units.

1. All the seven base units in the SI system are, however, not of *fundamental* significance. The question as to how many fundamental units are essential in the physical sciences, and which units are to be considered as fundamental, is a deep one. Indeed, it is related to the question as to which theory (or theories) constitutes a fundamental description of nature, and how many independent fundamental constants are included in that theory, such a constant in a theory being one which cannot be accounted for in terms of other constants at a more basic

level. In the early days of modern physics, Max Planck named three universal constants (Planck's constant (h), the velocity of light (c), and the universal constant of gravitation (G)) as being of fundamental significance and pointed out that these could be made use of in defining the three fundamental units of length, time, and mass. Considerations of a practical nature led to the adoption of units for length, mass, and time that differed from the ones based on h , c , and G , as suggested by Planck.

In the present state of our knowledge, the most fundamental theory for the description of physical phenomena is the *standard model* of elementary particles in which the fundamental constants h and c make their appearance, while the third constant G relates to the theory of gravitation, where the latter has not been satisfactorily integrated with the standard model. It is possible that a fundamental theory more comprehensive than the standard model may be constructed at some future date, when our conception of fundamental units will possibly get altered.

The question of fundamental units, in other words, remains an open one at a deeper level. For our present purpose, the units of length, time, and mass will be assumed to constitute the set of fundamental units.

2. Strictly speaking, the dimension of a quantity is defined in relation to the three fundamental units of length, mass and time, and not to the seven base units in the SI system. However, the unit of electric current is also made use of in determining the dimensions of various physical quantities, while the other base units are also referred to in some instances in deciding what the dimensions of given physical quantities are to be. Convention, expediency, and redundancy, all are there in the choice of the units in terms of which the dimensions of physical quantities are defined. At the same time, all these are done away with when only length, mass and time are used as the fundamental dimensions. This, however, is seemingly at odds with the SI system where the dimension of current is accorded a role analogous to that of length, mass, or time. This results in a loss of simplicity (as reflected, for instance, in the appearance of the dimensional quantities ϵ_0 and μ_0 (refer to sec. 1.4.4, and also to chapters 11, chap12) in relations involving electrical and magnetic quantities) and contact with presently known fundamental principles of physics. However, the SI system will be made use of in this book since

it enjoys wide acceptance for historical reasons.

1.4 SI units, and dimensions

1.4.1 The seven base units

The seven base units in the SI system are those of length (meter, m), time (second, s), mass (kilogram, kg), electric current (ampere, A), thermodynamic temperature (kelvin, K), amount of substance (mole, mol), and luminous intensity (candela, cd). Among these, the first four are, in a sense, more fundamental compared to the remaining three since the latter can be reduced to these four. These four are used to identify the dimension of any given physical quantity.

However, a commonly used practice is to make use of the unit of charge (the coulomb, C) instead of that of electric current in identifying the dimensions and units of various physical quantities of interest. For this latter purpose, appropriate relations among the physical quantities are made use of. As mentioned above, the remaining three base units (K, mol, cd) are sometimes invoked in specifying the dimensions of a number of quantities.

The definitions of the seven base units are based on an elaborate set of specifications so that there may be as little ambiguity as possible in these. The principles involved in these definitions are as follows.

In the following definitions of the base units and in subsequent sections of this chapter, you will find reference to terms and concepts that will be discussed at greater length later in the chapter or in later chapters of the book.

1. The unit of time, the second (s).

The second (s) is defined in relation to the *time period* characterizing a particular spectral line corresponding to the transition between two hyperfine levels (levels

corresponding to a very small difference in the energy) of the cesium-133 (Cs^{133}) atom, being 9 192 631 770 times the said time period.

2. The unit of length, the meter (m).

The meter (m) is defined in relation to the *distance traveled by light in vacuum in one second*, being $\frac{1}{299\,792\,458}$ times the said distance. This results in a unique assignment of value for the velocity of light in vacuum.

3. The unit of mass, the kilogram (kg).

The kilogram (kg) is defined as the mass of a cylindrical body made of platinum-iridium alloy, kept under standard conditions at the International Bureau of Weights and Measures in France, this body being referred to as the *international prototype of the kilogram*.

4. The unit of current (A).

Consider two infinitely long straight parallel wires, each of negligible cross-section, carrying identical currents of such magnitude that the force (see sec 12.8.2) on either wire due to the other is 2×10^{-7} newton per meter of its length, the unit of force in the SI system being the *newton* (N). This current is then defined as being 1 ampere (A).

5. The unit of temperature, the kelvin (K).

The *triple point of water* (see sec. 8.22.5) corresponds to a unique temperature. A temperature measuring $\frac{1}{273.16}$ times the temperature of the triple point of water is termed a kelvin (K), and serves as the unit of temperature (more precisely, of temperature difference) in the SI system.

6. The unit of amount of substance, the mole (mol).

The mole is the unit of the amount of a substance made of some particular type of 'elementary entities', which may be atoms, molecules, ions, electrons, other parti-

cles, or specified groups of such particles. It represents that amount of substance which contains as many elementary entities as there are atoms in 0.012 kg of carbon 12 (C^{12}).

7. The unit of luminous intensity, the candela (cd).

Consider a source that emits monochromatic radiation of frequency 5.40×10^{14} hertz and that has a radiant intensity of $\frac{1}{683}$ watt per steradian in any given direction. The luminous intensity of that source in the given direction is then defined to be a candela, the unit of luminous intensity.

The concept of luminous intensity relates to *photometry*, a subject you will not find discussed further in this book.

Problem 1-1

Work out the time in second for light to travel through 1m. Assuming the velocity of sound in air to be $343.2 \text{ m}\cdot\text{s}^{-1}$, calculate the time taken by sound to travel through the same distance.

Answer to Problem 1-1

(Light) $\frac{1}{299\,792\,458} \text{ s} = 3.336 \times 10^{-9} \text{ s}$ (approx); (Sound) $2.914 \times 10^{-3} \text{ s}$ (approx).

1.4.2 Dimensions related to units

On determining the unit of a physical quantity in the SI system, arrived at from the appropriate defining relation(s), it is found that the unit can be expressed in terms of the base units mentioned above in the form of a product, each term of a product representing one of the base units raised to some exponent or power which may be a positive or negative integer or a rational number. These exponents then determine the dimension of the quantity under consideration. The dimension is expressed in terms of the dimensions corresponding to the base units.

The dimension is expressed with an upper case symbol enclosed within brackets. Thus, the dimension of length, mass and time are denoted as [L], [M], and [T] respectively. The dimension of current and temperature, which are also used, are denoted by [I] and [Θ]. The number of dimensions in terms of which other dimensions are expressed is, to some extent, context-dependent.

As an example, consider the fact that velocity is a ratio of distance and time. Hence, the dimension of velocity can be expressed as $[LT^{-1}]$, corresponding to its unit being $\text{m}\cdot\text{s}^{-1}$. Here the exponents relating to length and time are respectively 1 and -1 , which determine the dimension of velocity. Again, the defining formula for force can be seen to be of the form $\frac{\text{mass}\times\text{distance}}{(\text{time interval})^2}$. Correspondingly, the dimension and unit of force are, respectively, $[LMT^{-2}]$ and $\text{m}\cdot\text{kg}\cdot\text{s}^{-2}$, the more commonly used name for the latter being the *newton* (N).

On referring to the definitions of *work* and *torque* (see chapter 3), one finds that both are defined in terms of a product of a force and a distance. This means that the dimensions of both of these quantities are the same, namely, $[ML^2T^{-2}]$. While the units of the two quantities are also the same, namely $\text{kg}\cdot\text{m}^2\cdot\text{s}^{-2}$ in the SI system, these are usually referred to by different names, viz., *joule* (J) for work, and $\text{N}\cdot\text{m}$ for torque. The unit for *energy* in the SI system is also the joule.

If one confines oneself to the use of SI units alone, then no added advantage is to be derived from the use of dimensions as compared to that of the corresponding units. The usefulness of dimensions in physics lies in the construction of *dimensionless quantities* that helps greatly in the setting up of instructive relations among various physical quantities, making possible, in particular, the analysis of physical situations in terms of the *principle of similarity*. The analysis of physical situations and of interrelations among physical quantities in terms of dimensions is broadly referred to as *dimensional analysis*. This I shall briefly turn to in section 1.6.

As an example of a dimensionless physical quantity, one can refer to the quantity termed *strain* describing the state of elastic deformation in a body. Since it is defined as a ratio

of two lengths (see sec. 6.3) it is a dimensionless quantity, and one does not need a unit to specify its value. Other dimensionless quantities can be constructed by making use of the relations among physical quantities of various definitions.

As already mentioned, I shall present examples of units and dimensions of a number of physical quantities in this chapter that will be introduced in greater details only in subsequent chapters of the book. You may, if you like, refer to these chapters in gaining a better understanding of these quantities but, in the main, all I want you to pick up from the present chapter is an idea as to how the units and dimensions of various quantities are to be arrived at from known relations among these, without necessarily getting to know the underlying concepts and derivations.

1.4.3 Derived units: selected physical quantities

As mentioned above, the SI system includes a list of twenty two special derived units for which specific names have been agreed upon. The units of all other physical quantities are expressed in terms of these twenty two special derived units and the seven base units. I indicate below the units and dimensions of a number of physical quantities we will encounter in later chapters in this book.

1. The unit and dimension of pressure and stress

Both the units of force and pressure belong to the list of twenty two special derived units, the former having already been introduced above as the newton (N). Since pressure is defined as force per unit area ($\frac{\text{force}}{\text{length}^2}$), the unit of pressure works out to $\text{N}\cdot\text{m}^{-2}$. This is given the name *pascal* (Pa). The dimension of force being $[\text{MLT}^{-2}]$, that of pressure is $[\text{ML}^{-1}\text{T}^{-2}]$.

The unit of *stress* (see sec. 6.4) at any point in a deformable body is also $\text{N}\cdot\text{m}^{-2}$. As a matter of fact, pressure is a special instance of stress.

2. The unit and dimension of power

The unit of energy has been introduced above as the joule (J), its dimension being $[ML^2T^{-2}]$. The unit and dimension of work are also the same. Power is defined as rate of doing work, $(\frac{\text{work}}{\text{time}})$, and hence its dimension is seen to be $[ML^2T^{-3}]$. The unit of power ($J \cdot s^{-1}$) is given the name *watt* (W). It also denotes the rate of energy output from a system, such as a source of sound or an optical source.

3. Frequency, angular frequency, and angular velocity

Frequency is defined in the form $\frac{1}{\text{time}}$. Accordingly, its dimension is $[T^{-1}]$. Its unit is given the special name *hertz* (Hz), being one among the twenty two special derived units in the SI system. From a fundamental point of view, the quantities angular frequency, and angular velocity have the dimension $[T^{-1}]$ as well. But the unit of either of these two quantities is referred to as $\text{rad} \cdot s^{-1}$, where the *rad* denotes the unit of plane angle in the SI system. Since the measure of a plane angle is defined in the form of a ratio of two lengths, it is a dimensionless quantity. When expressed in terms of the base units, it is sometimes written in the form $m \cdot m^{-1}$.

4. Surface tension and coefficient of viscosity

Surface tension is a property characterizing a liquid (or, more precisely, characterizing an *interface*; it is, however, commonly used to characterize the surface properties of a liquid; see chapter 7) and is defined in the form $\frac{\text{force}}{\text{length}}$. Accordingly, its dimension is $[MT^{-2}]$, and its unit is $N \cdot m^{-1}$.

Viscosity is a property characterizing a fluid (see section 7.5). The coefficient of viscosity is defined by a relation of the form

$$\text{force} = \text{coeff. viscosity} \times \text{area} \times \frac{\text{velocity}}{\text{distance}}.$$

Accordingly, the coefficient of viscosity is characterized by the dimension $[ML^{-1}T^{-1}]$, and its unit is Pa·s.

5. The unit and dimension of charge

In the SI system, the ampere, the unit of current, is chosen as one of the base units. All electrical and magnetic quantities relate in one way or other to this base unit. The definition of charge takes the form

$$\text{charge} = \text{current} \times \text{time},$$

though from a conceptual point of view, it is current that is more commonly defined in the form $\frac{\text{charge}}{\text{time}}$. Denoting the dimension of current by [I], the dimension of charge is seen to be [IT]. The unit of charge, (A·s), is given the special name *coulomb* (C).

6. Potential difference and electric field intensity

According to the definition of electric potential, energy and potential are related in the form

$$\text{energy} = \text{charge} \times \text{potential}.$$

Hence the dimension of potential is $[\text{ML}^2\text{T}^{-3}\text{I}^{-1}]$, and its unit in the SI system is $\text{J}\cdot\text{C}^{-1}$. It is given the special name *volt* (V).

The electric field intensity (or, in short, electric intensity, also referred to as electric field strength) is related to the potential in the form

$$\text{intensity} = \frac{\text{potential}}{\text{distance}}.$$

While the intensity is a vector quantity, its dimension and unit are determined regardless of its direction (see chapter 2), and the same dimension and unit hold for any Cartesian component of the intensity as well. Evidently, the dimension of electric field intensity is $[\text{MLT}^{-3}\text{I}^{-1}]$ and its unit is referred to in the form $\text{V}\cdot\text{m}^{-1}$, an alternative form (not commonly used) being $\text{N}\cdot\text{C}^{-1}$.

7. Magnetic flux density and magnetic field strength

There exists a considerable degree of non-uniformity (and some degree of confusion) relating to the names by which the magnetic field vectors are referred to.

The magnetic analog of the electric field intensity (or electric field strength) vector is commonly denoted by the symbol \mathbf{B} , and is defined with reference to the force exerted on a moving charged particle (the *Lorentz force*) or on a current-carrying conductor in the field (see sec. 12.8.4).

In this book we refer to this vector as the magnetic field intensity. Its definition corresponds to a relation of the form

$$\text{magnetic field intensity} = \frac{\text{force}}{(\text{current}) \times (\text{length})}.$$

This gives its unit as $\text{N} \cdot \text{m}^{-1} \cdot \text{A}^{-1}$, which is given the name *tesla* (T), it being one of the twenty two specially named derived units in the SI system. The dimension of this quantity is, accordingly, $[\text{MT}^{-2}\text{I}^{-1}]$.

What I have termed the magnetic field intensity here and elsewhere in this book is referred to as the magnetic *flux density* in the SI system. At times the name ‘magnetic intensity’ or ‘magnetic field strength’ is also used.

The source of the confusion lies in the fact that there exists *another* vector in magnetism, commonly denoted by the symbol \mathbf{H} , that is of relevance in describing magnetic fields in material media (the corresponding vector describing electrical fields in material media is termed the ‘electric displacement’, and is denoted by \mathbf{D}). In this book, we will have little occasion to work with this vector. It is commonly referred to as the magnetic field strength (any confusion with \mathbf{B} is usually settled by referring to the context). Its dimension is $[\text{L}^{-1}\text{I}]$, but its unit is commonly written as *amp-turn per meter*, though the simpler $\text{A} \cdot \text{m}^{-1}$ is also used.

Problem 1-2

The *kinematic viscosity* of a fluid is defined by an expression of the form $\frac{\text{coefficient of viscosity}}{\text{density}}$ (see sec. 7.5.2.1). Obtain its dimension.

Answer to Problem 1-2

$$[\text{L}^2\text{T}^{-1}].$$

Problem 1-3

Problem: The *energy of a magnetic dipole* (see sec. 12.8.12.1; the magnetic dipole is characterized by a dipole moment, a vector quantity, see sec. 1.7) in a magnetic field is given by an expression of the form

$$\text{energy} = \text{dipole moment} \times \text{magnetic field intensity}.$$

Obtain the dimension of magnetic dipole moment, and its unit in the SI system.

Answer to Problem 1-3

$$[\text{L}^2\text{I}], \text{ A} \cdot \text{m}^2.$$

Problem 1-4

The *capacitance* of a capacitor (see sec. 11.11) is defined by means of a formula of the form $\text{capacitance} = \frac{\text{charge}}{\text{potential difference}}$, and its unit in the SI system is named a *farad*. Express the farad in terms of the joule, the ampere, and the second.

Answer to Problem 1-4

$$\text{J}^{-1} \cdot \text{A}^2 \cdot \text{s}^2.$$

Problem 1-5

The *coefficient of self inductance* (see sec. 13.2.6) of an inductor is related to magnetic energy by a relation of the form $\text{energy} = \text{self inductance} \times \text{current}^2$. Obtain the unit of self inductance in terms of the joule and the ampere. It is referred to as the *henry* in the SI system.

Answer to Problem 1-5

$$\text{J} \cdot \text{A}^{-2}.$$

1.4.4 Units and dimensions of a few physical constants

1. The universal constant of gravitation.

The universal constant of gravitation (or, the gravitational constant, G) appears in the formula for the force of gravitational attraction between two massive particles, in accordance with Newton's law of gravitation (see section 5.1). According to this formula, the defining expression for G is of the form $\frac{\text{force} \times \text{distance}^2}{\text{mass}^2}$. This means that the dimension of G is $[M^{-1}L^3T^{-2}]$, and its unit is $\text{m}^3 \cdot \text{kg}^{-1} \cdot \text{s}^{-2}$ or, equivalently, $\text{N} \cdot \text{m}^2 \cdot \text{kg}^{-2}$. An approximate value of the universal constant of gravitation is

$$G = 6.673 \times 10^{-11} \text{ N} \cdot \text{m}^2 \cdot \text{kg}^{-2}. \quad (1-1)$$

2. Boltzmann's constant.

Boltzmann's constant (k_B) is an important constant in thermal physics as also in other related areas. Its defining expression is of the form $\frac{\text{energy}}{\text{temperature}}$ (in this context, see section 8.4.4). Accordingly, its dimension is $[ML^2T^{-2}\Theta^{-1}]$, where $[\Theta]$ stands for the dimension of temperature, and its unit is $\text{J} \cdot \text{K}^{-1}$. An approximate value of this constant is

$$k_B = 1.381 \times 10^{-23} \text{ J} \cdot \text{K}^{-1}. \quad (1-2)$$

3. The universal gas constant.

The universal gas constant (R) appears in the *equation of state* of an *ideal gas* (see section 8.10.3), and is relevant in describing the behaviour of a real gas as well. Its defining expression is of the form $\frac{\text{pressure} \times \text{volume}}{\text{mole number} \times \text{temperature}}$. In this expression, the numerator is a quantity with the dimension of energy. Hence the dimension of the gas constant is $[ML^2T^{-2}\Theta^{-1}\text{mol}^{-1}]$, and its unit is $\text{J} \cdot \text{K}^{-1} \cdot \text{mol}^{-1}$. An approximate

value of this constant is

$$R = 8.314 \text{ J} \cdot \text{K}^{-1} \cdot \text{mol}^{-1}. \quad (1-3)$$

4. Planck's constant.

Planck's constant (h) was introduced by Max Planck while explaining the energy distribution in *black body radiation* (see section 16.9) and marked the beginning of a new era in physics. It also features in the de Broglie relations (section 16-5) that served as the basis of quantum theory. It satisfies a relation of the form

$$\text{energy} = \text{Planck's constant} \times \text{frequency}.$$

Accordingly, its dimension is $[\text{ML}^2\text{T}^{-1}]$, and its unit is J·s. An approximate value of Planck's constant is

$$h = 6.626 \times 10^{-34} \text{ J} \cdot \text{s} \text{ (approx)}. \quad (1-4)$$

5. Stefan's constant.

Stefan's constant (σ) relates to the rate of radiation of energy from a black body (see section 8.23.3.1) and can be expressed in terms of Planck's constant and Boltzmann's constant. It satisfies a relation of the form

$$\frac{\text{energy}}{\text{time}} = \text{Stefan's constant} \times \text{area} \times (\text{temperature})^4.$$

Accordingly, its dimension is $[\text{MT}^{-3}\Theta^{-4}]$, and its unit is $\text{J} \cdot \text{s}^{-1} \cdot \text{m}^{-2} \cdot \text{K}^{-4}$. An approximate value of Stefan's constant is

$$\sigma = 5.67 \times 10^{-8} \text{ W} \cdot \text{m}^{-2} \cdot \text{K}^{-4}. \quad (1-5)$$

6. Permittivity and permeability of free space.

Electric and magnetic fields can be set up in free space by charges and currents respectively. The strengths of these fields for given charges and currents can be worked out with the help of formulae that involve two constants referred to as the permittivity and permeability of free space, these being denoted by symbols ϵ_0 and μ_0 respectively. These two constants, however, are not independent, and are related through the *velocity of light in vacuum*, another fundamental constant in physics.

The permittivity of free space can be expressed by a formula of the form $\frac{(\text{charge})^2}{\text{force} \times (\text{distance})^2}$. Accordingly, its dimension is $[M^{-1}L^{-3}T^4I^2]$, and its unit is $C^2 \cdot N^{-1} \cdot m^{-2}$. An approximate value of ϵ_0 is

$$\epsilon_0 = 8.85 \times 10^{-12} \text{ C}^2 \cdot \text{N}^{-1} \cdot \text{m}^{-2}. \quad (1-6)$$

The permeability of free space is similarly given by a formula of the form $\frac{\text{force}}{(\text{current})^2}$, from which its dimension is seen to be $[MLT^{-2}I^{-2}]$, and hence its unit is $N \cdot A^{-2}$. The value of this constant is

$$\mu_0 = 4\pi \times 10^{-7} \text{ N} \cdot \text{A}^{-2}. \quad (1-7)$$

The two constants ϵ_0 and μ_0 are related to each other as

$$\epsilon_0 \mu_0 = \frac{1}{c^2}, \quad (1-8)$$

where c stands for the velocity of light in vacuum (approx. $3 \times 10^8 \text{ m} \cdot \text{s}^{-1}$).

Problem 1-6

The Stefan constant is related to Planck's constant, Boltzmann's constant, and the velocity of light by a relation of the form

$$\sigma = \frac{2\pi^5 k_B^4}{15h^3 c^2},$$

where a stands for an integer exponent. Making use of known dimensions of h , k_B , and c , obtain the exponent a .

Answer to Problem 1-6

$a = 4$.

1.4.5 Prefixes denoting multiples and submultiples

Suppose that an electron travels a certain path in a time 1.1×10^{-6} s. One can express this more conveniently by saying that the time of travel is 1.1 microsecond (μ s), where the prefix *micro* is attached to the unit second (s) to denote a certain submultiple of a second, namely, 10^{-6} s. Similarly, when one says that the frequency of a certain microwave source is 30 gigahertz (GHz), one actually means a frequency of 30×10^9 Hz, since the prefix *giga* stands for a multiple of 10^9 . This approach of using prefixes to represent multiples and submultiples in powers of 10 happens to be a convenient one when the value of a physical quantity is large or small compared to the relevant unit. A few examples of prefixes are:

giga (10^9), mega (10^6), kilo (10^3), hecto (10^2), deca (10^1), deci (10^{-1}), centi (10^{-2}), milli (10^{-3}), micro (10^{-6}), nano (10^{-9}), pico (10^{-12}).

1.5 Other systems of units

1.5.1 Systems of units other than the SI system

The SI system is not the only possible system of units for representing the values of physical quantities. Other systems have proved useful in various different contexts, and several are still in use in different areas in the physical sciences.

For instance, the *cgs* system of units makes use of the centimeter (cm), gram (g), and the second (s) as units of length, mass, and time respectively. While these three are sufficient to consistently represent the units of physical quantities in mechanics, further considerations are needed when one broadens the area of discourse to include electrical

and magnetic quantities. For this, the cgs *electrostatic units* (esu) and the cgs *electromagnetic units* (emu) have proved to be useful alternative systems. Another system of great usefulness in the consideration of fundamental theories in physics is the *Gaussian* system, a still more elegant and convenient system in fundamental physics being the *natural* system of units.

1.5.2 Conversion from the SI to other systems of units

However, I shall not enter into details relating to these alternative systems of units. What is important to point out in this context is that, whatever system of units one chooses to use, one has to keep on *consistently* using the same system in making a derivation or a set of derivations in a chosen field. If, in the process, a physical quantity is encountered which is expressed in a system other than the system of units being used, it is to be *converted* appropriately.

For instance, the unit of force in the cgs system is $1 \text{ g}\cdot\text{cm}\cdot\text{s}^{-2}$, which is referred to as a *dyne*. Noting that $1 \text{ cm} = 10^{-2} \text{ m}$, and $1 \text{ g} = 10^{-3} \text{ kg}$, one gets the relation $1 \text{ dyne} = 10^{-5} \text{ N}$, which is the conversion formula for the unit of force from the cgs to the SI system. Similarly, the unit of energy, or work, in the cgs system is an *erg* which is equivalent to $1 \text{ dyne} \times 1 \text{ cm}$. Converting from dyne to newton and from centimeter to meter, one obtains $1 \text{ erg} = 10^{-5} \text{ N} \times 10^{-2} \text{ m} = 10^{-7} \text{ J}$, which constitutes the conversion formula for energy from the cgs to the SI system.

1.5.3 A few convenient non-SI units

There exist a number of areas of study in which specially defined units are used, these having been found convenient for such areas. Their continued use has precluded SI units, even with appropriate prefixes attached, from being made use of, except for the purpose of occasional reference. For instance, in atomic and nuclear physics, and in the physics of elementary particles, the unit of energy that is commonly used is the *electron volt*, along with appropriate prefixes, rather than the joule. The electron volt (abbreviated eV) is defined as the change in energy of an electron (which carries a charge of magnitude $1.6 \times 10^{-19} \text{ C}$) as it moves through a potential difference of 1 V, and converts

to 1.6×10^{-19} J in the SI system. While the eV is commonly used in atomic physics, a larger unit, the MeV ($=10^6$ eV) is more convenient for use in nuclear physics.

Another such specially defined unit is the unified *atomic mass unit* (symbol, u) or the *dalton* (Da), defined as $\frac{1}{12}$ times the mass of an isolated carbon-12 atom at rest and in the *ground state* (refer to sections 16.5.3, 18.4.1). It is used to specify nuclear, atomic and molecular masses, and converts to (approx.) 1.66×10^{-27} kg in the SI system. A larger unit, the kilodalton (10^3 Da), is used in the field of biochemistry and molecular biology, where large molecules are studied.

Yet another example of a specially defined unit is the *light year* (ly), which is the unit of distance on an astronomical scale. It is defined as the distance travelled by light in vacuum in one Julian year (the year being itself a specially defined non-SI unit). However, another unit, the *parsec* ($= 3.26$ ly (approx)) is more commonly used in the technical literature.

In engineering, the *horsepower* (HP) is often used as a convenient non-SI unit of power. However, there exist several definitions of the HP, which differ somewhat from one another and so, it is essential to clearly state which of these is being used in a given context. One of these, the unit of power rating for electric motors, converts to 746 W in the SI system.

Finally, in thermal physics, one sometimes uses the *calorie* as a unit of heat, and the *Celsius* as a unit of temperature (or, more precisely, *temperature difference*). The calorie, now being gradually replaced by the joule in thermal physics, converts to ≈ 4.2 J in the SI system. The Celsius scale of temperature is widely used today in meteorology. As an interval of temperature, 1 Celsius degree equals 1 kelvin degree on the SI scale. However, temperatures on the Celsius scale are related to those in the kelvin scale by a constant difference, namely,

$$t \text{ } ^\circ \text{C} = (t + 273.15) \text{ K.}$$

1.6 Dimensional analysis

There exists a vast network of relations among the physical quantities whose values can be determined by appropriate means. Any such relation is, in general, in the form of an equality, with each side of the equality expressed as a sum of terms (which may reduce to just one term as well), each term being, in turn, a product of several factors (or just one single factor). For instance, one encounters the following relation in mechanics for the motion of a particle along a straight line under a constant force:

$$s = ut + \frac{1}{2}at^2, \quad (1-9)$$

where s stands for a distance, u for a velocity, a for an acceleration, and t for a time interval. In this relation, there occurs the term $\frac{1}{2}at^2$, which is a product of several factors. Of these, the first factor ($\frac{1}{2}$) is a dimensionless number and the other two are physical quantities with appropriate dimensions where, moreover, the quantity t occurs with an exponent 2.

Looking at any term of this type occurring in any given relation among several physical quantities, one can determine its dimension. For instance, in the above example, the dimension of a is $[LT^{-2}]$, and that of t^2 is $[T^2]$, which tells us that the dimension of the term $\frac{1}{2}at^2$ is $[LT^{-2}] \times [T^2]$, i.e., $[L]$.

Indeed, it is in this manner that the dimensions of all the physical quantities in sections 1.4.3 and 1.4.4 above have been determined. A determination of the dimensions of the various terms in a mathematical relation involving physical quantities, can often be useful and instructive. This approach of calculating and making use of the dimensions of terms in a mathematical relation is referred to as *dimensional analysis*.

1.6.1 Principle of dimensional homogeneity

Dimensional analysis is based on the *principle of dimensional homogeneity*: for a mathematical relation among a number of physical quantities to be a valid one, the dimensions of all the terms occurring in the relation have to be the same.

This principle derives from the requirement that one cannot add up disparate physical quantities, i.e., ones with different dimensions. For instance, it does not make sense to add up a length with a mass. More generally, it is not possible to add or subtract two terms with dimensions, say, $[L^a M^b T^c]$ and $[L^p M^q T^r]$ unless the exponents are pairwise same, i.e., unless $a = p$, $b = q$, $r = c$. Evidently, a similar requirement can be formulated in terms of the *units* of the various different terms in the relation under consideration where a single system of units, like, say, the SI system, is used. This is because of the fact that the dimensions and units have to correspond to one another.

Suppose someone derives a certain mathematical relation between a set of physical quantities from theoretical considerations. A necessary condition for the relation to be a valid one is then obtained from the principle of dimensional homogeneity, i.e., expressing each term in the form of a product of the basic dimensions with certain exponents, these exponents have to be the same for all the terms so expressed. For instance, in the relation (1-9), each of the three terms may be seen to be of the dimension $[L]$, i.e., the relation satisfies the necessary condition to be a valid one.

However, just the principle of dimensional homogeneity cannot guarantee the validity of the relation under consideration, i.e., the principle does not provide a *sufficient* condition for the validity. For instance a relation of the form $s = ut + \frac{1}{3}at^2$ instead of (1-9) would be dimensionally homogeneous but would not be a valid relation between the displacement, velocity, acceleration, and time interval.

1.6.2 An application: Stokes' formula for viscous drag force

Dimensional analysis can be invoked not only to check for the validity of mathematical relationships between physical quantities but, under certain circumstances, to *guess* at the form of the relationship between a set of quantities.

For instance, consider a spherical body of radius r moving with a velocity v through a fluid whose coefficient of viscosity is η (see sec. 7.5 for an introduction to the concept of viscosity of a fluid). What will be the resistive force (commonly referred to as the *drag* force, see sec. 7.5.8.3) acting on the body due to the viscosity of the fluid?

A complete solution of this problem is provided by the principles of *fluid dynamics*, at least when the velocity of the body does not have too large a value. However, assuming that the drag force depends only on the parameters r , v , and η , a partial solution can be arrived at as follows.

Let us assume that the required force is of the form

$$F = kr^a \eta^b v^c, \quad (1-10)$$

where k is some dimensionless constant and where a , b , c are exponents to be determined. Let us now apply the principle of dimensional homogeneity to the proposed relation (1-10). Using the known dimensions of F , r , η , and v , one finds that $[MLT^{-2}] = [L^a][M^b L^{-b} T^{-b}][L^c T^{-c}]$, which leads to the three equations

$$b = 1, \quad a - b + c = 1, \quad b + c = 2, \quad (1-11a)$$

from which the exponents a , b , c work out to

$$a = b = c = 1. \quad (1-11b)$$

This gives the required relation in the form

$$F = k\eta r v. \quad (1-12a)$$

Here the constant k cannot be evaluated by dimensional analysis since it is dimensionless. The more complete solution of the problem in fluid dynamics gives $k = 6\pi$. In other words, the complete expression for the drag force works out to

$$F = 6\pi\eta r v. \quad (1-12b)$$

This example indicates another limitation of dimensional analysis - one has to guess by some means the factors on which the force F can depend. For instance, the force might possibly depend on the density (ρ) of the fluid. One could then start from a proposed

relation of the form $F = kr^a\eta^bv^c\rho^d$ instead of eq. (1-10), with undetermined exponents a, b, c, d . Then, proceeding as above, one would have had three equations involving four unknowns, and a unique solution could not be obtained. In reality, however, the required relation does not involve the density of the fluid and dimensional analysis does help in finding a partial solution to the problem.

1.6.3 The principle of similarity

A mathematical relation between a number of physical quantities can sometimes be written in a *dimensionless* form. The physical quantities occurring in such a relation are first expressed in dimensionless form without altering their significance, and the relations are then written in terms of these dimensionless variables.

For instance, consider the pressure (p), molar volume (i.e., the volume per mole, v), and the temperature (T) of a gas. Such a relation is referred to as an equation of state, and several such equations, each of approximate validity, are known. An example of an equation of state is the Van der Waals equation which, written for one mole of the gas, is of the form:

$$(p + \frac{a}{v^2})(v - b) = RT. \quad (1-13)$$

Here R stands for the universal gas constant, and a and b are two constants whose values differ from one gas to another, the units of these constants being, respectively, $\text{J}\cdot\text{m}^3\cdot\text{mol}^{-2}$ and $\text{m}^3\cdot\text{mol}^{-1}$. One can now make use of the fact that each gas is characterized by a certain critical pressure (p_c), critical molar volume (v_c), and critical temperature (T_c), in terms of which the variables p , v , and T can be re-expressed in the form of dimensionless variables (referred to as *reduced variables*) as

$$x = \frac{p}{p_c}, \quad y = \frac{v}{v_c}, \quad z = \frac{T}{T_c}. \quad (1-14)$$

Interestingly, the equation of state, written with these dimensionless state variables of

a gas assumes the form

$$(x + \frac{3}{y^2})(y - \frac{1}{3}) = \frac{8}{3}z, \quad (1-15)$$

where one notes that the equation contains *no* constant, parameter, or variable that may differ from one gas to another, and is referred to as the *reduced equation of state*. Hence, when described in terms of the reduced variables, *all gases behave similarly*. More concretely, considering two different gases whose states are such that two of the variables x , y , and z have the same values for the two gases (say, $x_1 = x_2$, $y_1 = y_2$, where the suffixes '1' and '2' refer to the two gases respectively), then the third variable must also have the same value (i.e., $z_1 = z_2$). Evidently, the use of the state variables p , v and T would not have led to a similar conclusion because of the presence of the constants a , b in equation (1-13).

This shows that, when appropriate dimensionless variables are made use of, different systems (all belonging to a common class) behave identically since the mathematical relation expressing their behaviour assumes a *universal* form. This is what I mean by the term *principle of similarity*.

At times, a mathematical relation between physical quantities for a system or a class of systems contains, apart from constants that may differ from one system to another belonging to the class (like the constants a and b above), one or more *parameters* whose values depend on the condition under which the behaviour of a system is described. For instance, the equation of motion of a fluid through a pipe in the absence of heat transfer, when expressed in terms of dimensionless variables, involves a dimensionless parameter termed the *Reynolds number* ($R = \frac{\rho v D}{\eta}$, where D stands for the diameter of the pipe; other variables have been introduced above) which specifies in terms of a single number the physical condition under which the flow occurs.

This single dimensionless parameter then describes the characteristics of the flow regardless of the values of the relevant parameters (ρ , v , D , η) considered individually. For instance, one may consider the flows of two different fluids with different velocities

through pipes of different diameters. If, however, the Reynolds numbers be the same in the two cases, then the flows will be of a similar nature. If, for instance, the flow be a *turbulent* one (i.e., a flow with random fluctuations, which differs from a regular or *laminar* flow) for one of the two fluids, it will be turbulent for the other fluid as well (see sec. 7.5.9).

Digression: Buckingham Pi theorem.

In a paper written in 1915 in an informal style (which subsequently turned out to be an influential one), Rayleigh [1] pointed out the importance of the principle of similarity, which he referred to as 'the great principle of similitude'. In an equally influential paper [2] written in 1914, Buckingham gave a rigorous formulation of the principle in the form of what later came to be known as the *Buckingham pi theorem*. The theorem essentially tells us that if *all* the parameters and variables (including dimensional constants) relevant for the solution of a physical problem are known, then it can be formulated in relatively simple terms by introducing an appropriate set of dimensionless parameters and variables before actually attempting a solution. While the choice of the dimensionless variables (Buckingham denoted these with a common symbol 'pi' (the Greek letter), whence the name of his eponymous theorem) is a matter of insight into the problem, one also needs a correct formal procedure to set these up, which he outlined (refer to [3], chapter 1, for an exposition).

A restricted form of the theorem tells us how to effect a simplification in a relation among physical quantities by expressing it in a dimensionless form, reducing the number of variables (combinations of the variables appearing in the original relation) appearing in the simplified relation (see [4], chapter 7, for a more detailed exposition).

Feynman, in his inimitable style, gave an example of the application of the principle of similarity in [5], chapter 41. For a general introduction to principles of dimensional analysis, including a reference to the Buckingham pi theorem, you may have a look at [6].

If some of the variables relevant to a problem are not known or specified, then the efficacy of the principle of similarity is greatly reduced since it leads to only an incomplete formulation in terms of dimensionless variables, because dimensional variables continue to remain in the formulation.

The principle of similarity is a remarkably useful one in analyzing *complex* problems in the physi-

cal sciences and in engineering, where one does not hope for a complete solution but nevertheless wants to have as much insight as possible into the relation between the various parameters and variables defining the problem. The fact that a single dimensionless parameter, namely the Reynold's number, gives a universal characterization of a class of fluid flow problems, constitutes a nice illustration of this statement.

Problem 1-7

The *energy of deformation* per unit volume of an elastic body is of the form $kS^a\epsilon$, where k is a constant, S stands for the stress, ϵ for the strain, which is a dimensionless quantity (see sec. 6.3), and a is an undetermined exponent. Obtain the value of a by dimensional analysis.

Answer to Problem 1-7

$a = 1$.

Problem 1-8

The specific thermal capacity of a substance is defined by an expression of the form $\frac{\text{energy}}{\text{mass} \times \text{temperature interval}}$, while the thermal conductivity is defined by an expression of the form $\frac{\text{energy}}{\text{time} \times \text{length} \times \text{temperature interval}}$.

Show that the *Prandtl number* of a fluid, defined by an expression of the form $\frac{\text{specific thermal capacity} \times \text{coeff. of viscosity}}{\text{thermal conductivity}}$ is a dimensionless quantity. It is a convenient parameter for setting up the equation of heat flow in a moving fluid and for bringing out similarities in the flow features in various different situations.

Answer to Problem 1-8

HINT: Work out the dimension by making use of the dimensions of each of the quantities in the defining expression given, and check that all the exponents cancel.

1.7 Physical quantities as scalars and vectors

The dimension and unit of a physical quantity tells us something of its identity from a conceptual point of view - the dimension (and unit) differs for quantities whose conceptual definitions differ from one another or, more concretely, if the quantities under

consideration are related to other physical quantities in a different way and if, moreover, the measurement procedures for these differ distinctively from one another.

Incidentally, physical quantities can be classified into a number of different categories from *another* point of view as well - one that relates to a greater degree to the *mathematical* structure among the set of possible values of any given physical quantity. For instance, consider the possible values of the *volume* of a gas. Any such value can be specified by a *single* real number (which, in this instance, happens to be positive). On the other hand, the *velocity* of a particle cannot be specified so simply, since one requires a set of *three* components so as to specify the velocity completely. An equivalent way to say this is to state that one requires not only the magnitude, but the *direction* as well in order that the velocity is known completely.

To be sure, this difference between the two quantities, namely, volume and velocity, stems from the fact that they are distinct physical concepts, relating to other concepts in distinctive manners (which is why they correspond to different dimensions and units), but at the same time the distinction points to a difference in the manner their possible values are to be described mathematically.

More concretely, recall that the complete specification of the value of a physical quantity involves both a *numerical value* and a *unit*. It is in respect of the former, rather than the latter, that the distinction that I am now referring to arises. There are physical quantities like, say, volume and mass, whose numerical values can be specified with just one single real number. On the other hand, there are others like, velocity and force, each of which requires three real numbers for a complete specification. These are referred to as quantities belonging to the category of *scalars* and *vectors* respectively.

The next chapter will contain more detailed considerations relating to the concept of vectors as distinct from that of scalars since what I have said in these few paragraphs above does happen to be rather sketchy. But what I have to mention now is that there are other chapters to the story as well - the categorization of physical quantities does not stop at classifying them into the categories of just scalars and vectors. There exist

other physical quantities of a more complex nature (complex, that is, from the point of view of the mathematical description of their possible numerical values), referred to as *tensors*.

For instance, you will see in chapter 6 that the state of strain at a point in a deformable body cannot, in general, be specified simply by one single, or even three, real numbers. Indeed, it requires a set of six real numbers to completely specify the state of strain at a point. Each of these six is, in its own right, a scalar. But the six together characterizes a single physical condition, namely, the state of strain at a point in the body. What one states in technical language is that the state of strain at a point is a *symmetric tensor of rank two*.

This brings me to another interesting aspect of the story, namely, that tensors have their own classification scheme in terms of what is referred to as their *rank*. Indeed, from a broader point of view, scalars and vectors can also be looked upon as tensors - ones of rank zero and one respectively.

Scalars, vectors, and tensors will also be encountered in chapter 17 in the context of classifying physical quantities relating to space-time events in the special and the general theories of relativity.

At times, it serves some purpose to be acquainted with a concept even at the level of just getting to know the names of a few ideas underlying the concept. Because it helps you in orienting yourself when subsequently you set about learning of the concept in greater details. So, I will now close this section with just this short *summary*: physical quantities can be classified into tensors of various ranks from the point of view of the mathematical description of their possible numerical values, where *scalars* and *vectors* correspond to distinct categories of tensors of rank zero and one respectively.

Disclaimer: parsimony in citing references, and in posing exercises.

In this book, I will not inundate or overwhelm you with a lot of citations and references, because I harbor a secret aversion to references. Take, for instance, the references cited

in sec. intro-sec19 which I mentioned just for the sake of completeness. It will do you no harm in following the contents of this book even if you choose not to look these up now, before going on with the book (if, on the other hand, you do choose to have even a cursory look at those, now or later, your ideas will be broadened and enriched, and you may even be in a position to try out new ideas of your own). The same will hold for other citations in later chapters (there are only a very few of those altogether), where some of those will be included in the sub-text printed in small font.

The sub-text itself will constitute an important, though disposable, part of the book. You can disregard it as you move along, or you can selectively go through some of it, in which case you will still be able to follow the content without much impairment. However, reading the sub-text will often help you getting the bigger picture, and finding your way to things beyond the reach of this book. At some places, it contains notes and sidelights that you may find useful and interesting. You can read some, ignore some, and decide to come back to some others at a later time.

Finally, a word on exercises and problems. This book does *not* intend to challenge your mind or to provoke you to test your own intelligence by way of posing a lot of tough problems for you to solve. There are lots of places where you can find such problems, really tough ones at that, and it will be your own decision whether or not you want to have your intelligence challenged. As for me, I have had endless hours of squirming under the glare of tough problems posed in text-books. My suggestion would be, go in for those tough ones, but at your own good time. Meanwhile, just read along, and get a feeling of how the basic concepts of physics hang together and make up a coherent and beautiful whole. As regards the relatively small number of problems I will include in the various chapters of this book, there will be hints for the solution of most of those, or even complete solutions if need be. Some of the problems are 'elementary' and some 'advanced'. Take your pick, don't feel constrained.

Chapter 2

Vectors

2.1 Introduction

In this chapter I will introduce a certain class of physical quantities that require magnitudes as well as *directions* for their complete description. These can be represented in terms of appropriate *directed line segments*, and are termed *vectors*. To be precise, however, just a directed line segment cannot represent a physical quantity fully, because one needs to specify the *unit* of that physical quantity as well. Thus, one has to understand from the context whether the term vector refers to just a mathematical entity, without any unit attached to it, or to a physical quantity with its appropriate unit. In the present chapter I shall use the term mostly in the first of these two senses, i.e., as a mathematical entity that can be represented by a directed line segment.

The length of the directed line segment will denote the *magnitude* of the vector, while its direction will give us the *direction* in space of the vector. Such an entity will be depicted pictorially with the help of an arrow of appropriate length as in fig. 2-1.

I should point out here that in mathematics the term vector has a *broad*er connotation.

In this broader point of view, there may be various different *types* of vectors, each type being defined in terms of a certain set termed a *linear vector space*. In other

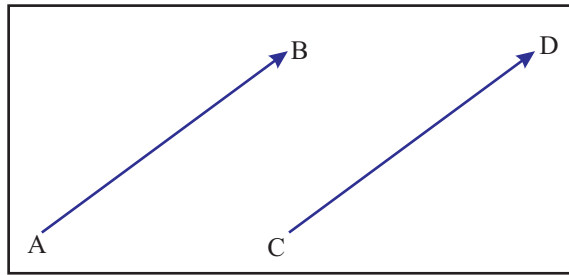


Figure 2-1: A directed line segment as a vector; the length of the segment represents the magnitude of the vector while its direction in space gives the direction of the vector; the segment extending from A to B stands for the same vector as the one extending from C to D because both the segments have the same length and also point in the same direction.

words, various different linear vector spaces define different classes or types of vectors. The directed line segments in the three dimensional physical space around us then constitute just one *instance* of vectors, making up a linear vector space of their own.

In order to qualify as a linear vector space a set has to satisfy certain requirements which I will not go into here. The principal characteristic underlying all such sets is that of *linearity*. This will become clearer from what I shall have to say in the following sections.

Each linear vector space is characterized by a certain *dimension*. I have mentioned above the three dimensional physical space we live in, which relates to the vector space of all possible directed line segments - a vector space of dimension three. On the other hand, the set of all directed line segments lying in any given plane (or in planes parallel to it) forms a vector space of dimension two.

Every linear vectrospace has, associated with it, a set of a special type, called a *field*. The elements of this field are termed *scalars* with reference to the vector space under consideration. A familiar example of a field is the set of *real numbers*, while the *complex numbers* constitute another example of a field. Recall that a complex number can be expressed in the form $a + ib$ where a and b are any two real numbers and the symbol i stands for the imaginary number $\sqrt{-1}$. However, while talking of the linear vector space of directed line segments - whose elements we will refer to as vectors in this book - it is the field of real numbers that is to be considered as the set of associated

scalars. A physical quantity that can be described with the help of just one single real number (along with an appropriate unit) is commonly termed a scalar. This usage of the term scalar is distinct from the definition that a scalar is a member of the field associated with a vector space, though it is closely related with the above statement that the set of real numbers constitutes the field of scalars associated with the vector space of directed line segments.

In brief, then, a vector in this book will stand for a directed line segment while a scalar will mean a real number. I will introduce below a *binary operation* among the set of vectors, to be referred to as *vector addition*. Another operation, to be called *multiplication of a vector with a scalar*, will also be introduced. Together, these two operations can be made use of in order to define, in a general way, a linear vector space.

2.1.1 Equality of two vectors

I must tell you at the outset that two directed line segments of the same length and parallel to each other i.e., ones pointing in the same direction, actually stand for the *same* vector. In other words, while representing a vector with a directed line segment, one needs to refer only to its length and its direction, but *not to its location in space*. Thus, in fig. 2-1, the line segment extending from A to B represents the *same* vector as the one extending from C to D. In other words, if one directed line segment can be made to coincide with another through *parallel translation*, then the two are to be considered identical from the point of view of representing a vector. Put differently, the directed line segment is described by specifying an *initial* point and a *final* point (the direction of the vector being from the initial to the final point), but a translation of both the initial and final points by any given parallel transport of the associated line segment does not alter the vector under consideration.

Thus, any given directed line segment corresponds to a unique vector, but any given vector corresponds to an infinite number of directed line segments (reason this out), each of which can be said to *represent* the given vector.

Problem 2-1

Imagine two vectors in a plane, one stretching from the point A (1, 3) to point B (5, -7), and the other from the point P to Q (-3, 4) where the Cartesian co-ordinates (with reference to a specified pair of axes) of points are indicated with the numbers within brackets. What should be the co-ordinates of the point P for the two vectors to be equal?

Answer to Problem 2-1

(-7, 14); you may skip this problem and read along till you have a look at sec. 2.5.1.

Vectors are usually represented symbolically either with letters of the alphabet with arrows overhead (like, for instance, \vec{A} , \vec{B}) or with the help of boldface letters (**A**, **B**). Alternatively, a vector represented by a directed line segment extending from, say A to B, can be named \vec{AB} . Thus, if the directed line segment AB, representing a vector, say, **A**, can be made to coincide with another directed line segment CD, representing **B**, through parallel translation (fig. 2-1) then one has

$$\mathbf{A} = \mathbf{B}. \quad (2-1)$$

2.1.2 Magnitude of a vector

Looking at the directed line segment AB of fig. 2-1 representing the vector **A**(say), the magnitude or *norm* of the vector is represented by the length of the segment, and is expressed as $|\mathbf{A}|$ (or, alternatively, as $\|\mathbf{A}\|$; at times one simply writes this as *A* for the sake of brevity). Similarly, the norm of **B** is expressed as $|\mathbf{B}|$ (or, simply, as *B*). Evidently, the norm of a vector is necessarily a non-negative quantity, i.e., for any vector **A**,

$$|\mathbf{A}| \geq 0. \quad (2-2)$$

The norm is, in general, *positive*. It can be zero, but only if the vector itself is the *zero vector*, which we now turn to.

2.1.3 The null vector

Imagine that the length of a directed line segment is being reduced till, in the end, the length becomes zero. This may then be termed a directed line segment of length zero and is referred to as the *zero* or *null* vector. Notice that the direction of the null vector is *indeterminate*, because whatever be the direction of the line segment to start with, in the end it is reduced to a point, when there remains no trace of the direction. In other words, the null vector is a vector of zero norm, having indeterminate direction. It may be denoted by the boldface symbol $\mathbf{0}$, but more commonly one uses the symbol 0 since it can be usually seen from the context that it is the null vector that is being referred to.

In sec. 2.5 we will get introduced to the concept of Cartesian components of a vector with reference to any specified Cartesian co-ordinate system. The null vector is one whose components are $(0, 0, 0)$ with reference to any arbitrarily specified Cartesian system.

2.2 Operations with vectors

2.2.1 Addition of vectors

2.2.1.1 Addition of two vectors

Figure 2-2 shows two vectors A and B represented by directed line segments OA and OB where both the segments have been taken to have the same initial point (O) (recall that it is only the length and direction of a segment that matters, and not the initial or final point). Imagine now a parallelogram $OABC$, with OA and OB as adjacent sides. Look at the diagonal segment OC of this parallelogram, with O as one end-point. Then the vector \vec{OC} , represented by the directed line segment OC , is termed the *vector sum* (or, in brief, *sum*) of the vectors A and B . Using the name C for this vector, one writes

$$A + B = C. \quad (2-3)$$

Notice that, given *any* two vectors, one can form their sum in this way to obtain a vector

as the result. In other words, the process of forming the vector sum is a *binary operation* within the set of vectors. The above geometrical approach of forming the vector sum is sometimes referred to as the *parallelogram rule*.

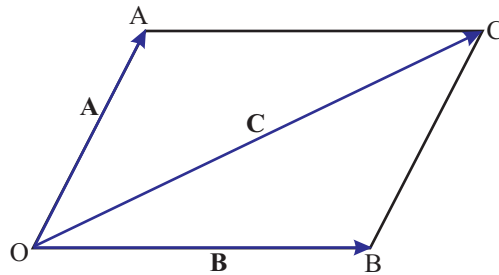


Figure 2-2: Addition of two vectors, the parallelogram rule; C is the vector sum of A and B.

2.2.1.2 Addition of more than two vectors

The definition of vector addition can be extended in a straightforward way to a set of more than two vectors. For instance, for any three vectors A, B, C, one can define the vector $A + (B + C)$ by first forming the sum of the vectors B and C, and then taking the sum of A with the resulting vector. At times, the sum of two or more given vectors is referred to as their *resultant*.

Problem 2-2

Referring to fig. 2-2, if the magnitudes of the vectors A and B be 5 and 3 respectively, and if the angle between them be $\frac{\pi}{3}$, work out the magnitude of their resultant C.

Answer to Problem 2-2

HINT: From the geometry of the figure, $|C|^2 = |A|^2 + |B|^2 + 2|A||B|\cos \frac{\pi}{3}$, i.e., $|C| = 7$.

By a consistent application of the parallelogram rule, one can check that $A + (B + C) = (A + B) + C$ (check this out), so that one can write either of the two expressions as

$A + B + C$. Further properties of the operation of vector addition also follow from its definition (see sec. 2.2.3).

The operation of addition of two or more vectors can be described in an alternative way as well. Suppose that two vectors A and B are to be added up and that the directed line segments AB and BC (fig. 2-3(A)) represent these two vectors, where the line segments are so placed that the initial point (the 'tail') of one of the two coincides with the final point (the 'head') of the other. The directed line segment (AC in the figure) extending from the tail of the latter to the head of the former will then represent the sum (C) of the two given vectors.

The prescription can be extended to the addition of more than two vectors, as illustrated in fig. 2-3(B), where the addition of three vectors is shown and where one has $A + B + C = D$. It may be mentioned here that, in the case of addition of two vectors, the directed line segments representing the two necessarily lie in a plane when placed in the manner mentioned above while, in the case of more than two vectors, the segments may not lie in a single plane.

2.2.2 Multiplication of a vector with a scalar

Look at fig. 2-4, where the vector A is represented by the directed line segment OA , and assume that p is any scalar, i.e., a real number. Now imagine a directed line segment OB whose direction is the same as that of OA , but whose length is p times that of OA . Then this directed line segment represents a vector termed the scalar multiple of A by the scalar p , which we write as pA , read as p times A).

If p happens to be a negative number, then pA is represented by a directed line segment of length $p|A|$ parallel to that representing A , but pointing in the *opposite* direction.

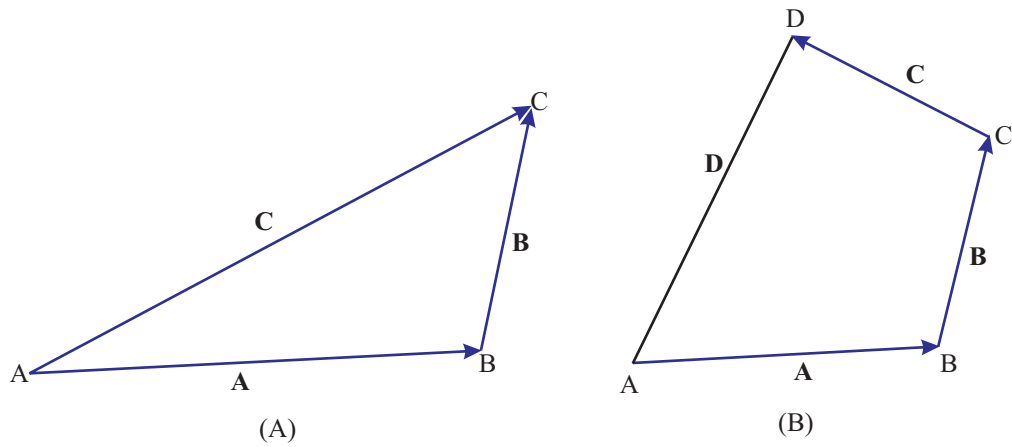


Figure 2-3: Illustrating the addition of two or more vectors; (A) two vectors \vec{A} and \vec{B} are represented by directed line segments AB and BC where the initial point of BC is made to coincide (by parallel translation) with the final point of AB ; the directed line segment AC extending from the initial point of AB to the final point of BC will then represent \vec{C} , the vector sum of the two vectors; (B) sum of more than two vectors; the directed line segments representing vectors \vec{A} , \vec{B} , \vec{C} are placed with the 'head' of one segment coinciding with the 'tail' of the next; the sum \vec{D} is then represented by the directed line segment extending from the tail of the first to the head of the last segment; for more than two vectors, the directed line segments need not be co-planar.

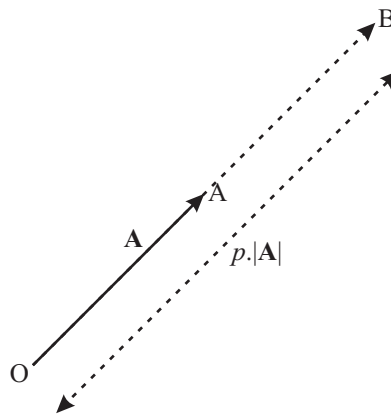


Figure 2-4: Illustrating multiplication of a vector with a scalar; the length of OB is p times that of OA ; the directed line segments OA and OB represent respectively \vec{A} and $p\vec{A}$.

2.2.3 Features of vector addition and scalar multiplication

The following conclusions emerge from the definitions of the two operations introduced above.

1. If \vec{A} is any vector then multiplying it with the scalar unity (1) gives back \vec{A} , and multiplying with the scalar zero (0) gives the null vector. Further, the vector sum

of the null vector and any other vector, say, \mathbf{A} , gives back \mathbf{A} .

$$1 \cdot \mathbf{A} = \mathbf{A}, \quad (2-4a)$$

$$0 \cdot \mathbf{A} = 0, \quad (2-4b)$$

$$0 + \mathbf{A} = \mathbf{A} \quad (2-4c)$$

where the last relation means that the null vector acts as the *identity element* in vector addition.

2. If \mathbf{A} and \mathbf{B} be any two vectors, then

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}, \quad (2-5)$$

i.e., in other words, the binary operation of vector addition is *commutative*.

3. If \mathbf{A} , \mathbf{B} , and \mathbf{C} be any three vectors, then

$$\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}, \quad (2-6)$$

which means that vector addition is an *associative* binary operation. Because of the associativity property, one can write either side of (2-6) as $\mathbf{A} + \mathbf{B} + \mathbf{C}$ and, in this manner, construct the vector sum of any given number of vectors.

4. If p and q be any two scalars and \mathbf{A} , \mathbf{B} be any two vectors then

$$(p + q)\mathbf{A} = p\mathbf{A} + q\mathbf{A}, \quad (2-7a)$$

$$p(\mathbf{A} + \mathbf{B}) = p\mathbf{A} + p\mathbf{B}, \quad (2-7b)$$

$$(pq)\mathbf{A} = p(q\mathbf{A}). \quad (2-7c)$$

These relations express a number of *distributivity* properties of vector addition in relation to multiplication with a scalar.

5. If \mathbf{A} be any vector then the vector $(-1) \cdot \mathbf{A}$ acts as the reciprocal vector of \mathbf{A} with respect to vector addition, i.e., its vector sum with \mathbf{A} gives the null vector. We write this vector as $-\mathbf{A}$ (similarly the vector $(-p) \cdot \mathbf{A}$ is written as $-p\mathbf{A}$). Thus,

$$\mathbf{A} + (-\mathbf{A}) = 0. \quad (2-8)$$

The sum of the vectors, say, \mathbf{A} and $-\mathbf{B}$ is written, for the sake of brevity, as $\mathbf{A} - \mathbf{B}$. In this connection, note that an equation of the form

$$\mathbf{A} + \mathbf{B} = \mathbf{C}, \quad (2-9a)$$

implies

$$\mathbf{A} = \mathbf{C} - \mathbf{B}. \quad (2-9b)$$

In a broader context, the above features are made use of in *defining* a linear vector space in mathematics. As I have already mentioned, we shall use the term vector in this book to denote a directed line segment (where any two parallel line segments of equal lengths are identified with each other). However, in a broader context, the set of directed line segments (with the above identification) is just an instance of a linear vector space in this more general sense.

2.3 Unit vector

Look at the directed line segment OA in fig. 2-5, representing the vector, say, \mathbf{A} . Let P be a point on the segment OA such that the segment OP is of unit length. The vector

\vec{OP} is then a vector in the same direction as \vec{OA} ($= \mathbf{A}$), having unit norm, and is termed the *unit vector* along \mathbf{A} . We shall denote it by the symbol \hat{A} , i.e.,

$$\vec{OP} = \hat{A}. \quad (2-10)$$

Now, the length of the segment OA is, by definition, the norm of \mathbf{A} , i.e., $|\mathbf{A}|$. Thus, the length of OA is the same as that of OP , multiplied with $|\mathbf{A}|$. In other words, the above definition of the unit vector \hat{A} and that of multiplication with a scalar tells us that

$$\mathbf{A} = |\mathbf{A}| \cdot \hat{A}, \quad (2-11a)$$

or,

$$\hat{A} = \frac{1}{|\mathbf{A}|} \mathbf{A}. \quad (2-11b)$$

Evidently, every vector *other than* the null vector has a unique unit vector associated with it.

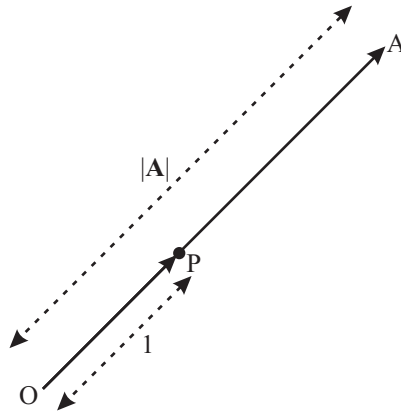


Figure 2-5: Illustrating the definition of a unit vector; the directed line segment OP represents the unit vector along the vector \mathbf{A} , represented by the directed line segment OA .

2.4 Scalar product of two vectors

The directed line segments OA and OB in fig. 2-6 represent the vectors, say, \mathbf{A} and \mathbf{B} . If θ be the angle between the two segments then the scalar $|\mathbf{A}||\mathbf{B}|\cos\theta$ is termed the *scalar product* (the term ‘dot product’ is also used) of the two vectors \mathbf{A} and \mathbf{B} . It is denoted by the symbol $\mathbf{A} \cdot \mathbf{B}$:

$$\mathbf{A} \cdot \mathbf{B} = |\mathbf{A}||\mathbf{B}|\cos\theta. \quad (2-12)$$

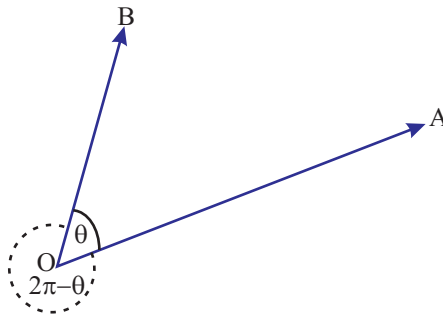


Figure 2-6: Illustrating the definition of scalar product of two vectors.

1. The product of two scalars is also expressed by inserting a dot in between the symbols for these scalars like, say, $p \cdot q$, but this dot is often left implied (one writes, simply, pq). However, it is necessary to use the dot in the left hand side of (2-12), which cannot be left implied.
2. Notice in fig. 2-6 that one can talk of *two* angles between the directed line segments OA and OB. The *smaller* of these two has been termed θ above. The other angle between the two segments is $2\pi - \theta$. However, so far as the scalar product of two vectors is concerned, it is immaterial as to which of the two angles is being called θ .

2.4.1 Features of scalar product

1. The scalar product of a vector with *itself* is nothing but the square of its norm because, in this special case, one has $\theta = 0$:

$$\mathbf{A} \cdot \mathbf{A} = |\mathbf{A}|^2. \quad (2-13)$$

As a result, the quantity $\mathbf{A} \cdot \mathbf{A}$ has to be either positive or zero and, moreover, it can be zero only when \mathbf{A} itself is the null vector. In other words, unless \mathbf{A} is the null vector, $\mathbf{A} \cdot \mathbf{A}$ is necessarily positive.

2. If \mathbf{A} and \mathbf{B} be any two vectors then

$$\mathbf{A} \cdot \mathbf{B} = \mathbf{B} \cdot \mathbf{A}, \quad (2-14)$$

i.e., in other words, the order of the factors involved in the scalar product can be interchanged without altering its value (*commutativity*).

3. If \mathbf{A} , \mathbf{B} , \mathbf{C} be any three vectors and p , q be any two scalars, then

$$\mathbf{A} \cdot (p\mathbf{B} + q\mathbf{C}) = p\mathbf{A} \cdot \mathbf{B} + q\mathbf{A} \cdot \mathbf{C}. \quad (2-15)$$

In a broader context, the above features are made use of so as to *define* a scalar product in a linear vector space. And then, the norm is defined with the help of (2-13). Further, orthogonality of two vectors (see below) is also defined in terms of this more general concept of scalar product.

4. If the directed line segments representing the vectors \mathbf{A} and \mathbf{B} are oriented perpendicularly to each other then

$$\mathbf{A} \cdot \mathbf{B} = 0, \quad (2-16)$$

because in this case one has $\theta = \frac{\pi}{2}$ in (2-12). The vectors \mathbf{A} and \mathbf{B} are then said to be *orthogonal* to each other.

2.4.2 Orthonormal triads of vectors

If three vectors are pairwise orthogonal to one another, then they are said to constitute an orthogonal triad. In particular, if each of the three vectors happens to be of unit norm, the triad is termed *orthonormal*. Thus, in fig. 2-7 (A), the directed line segments OA, OB, OC are each of unit length, and are, moreover, pairwise perpendicular to one another. Then, naming the corresponding vectors respectively \hat{i} , \hat{j} , and \hat{k} (notice the use of the caret symbol because all three are unit vectors here, refer to eq. (2-10)), we have

$$\begin{aligned}\hat{i} \cdot \hat{j} &= 0, \\ \hat{j} \cdot \hat{k} &= 0, \\ \hat{k} \cdot \hat{i} &= 0,\end{aligned}\tag{2-17}$$

and one says that the three unit vectors \hat{i} , \hat{j} , and \hat{k} constitute an orthonormal triad.

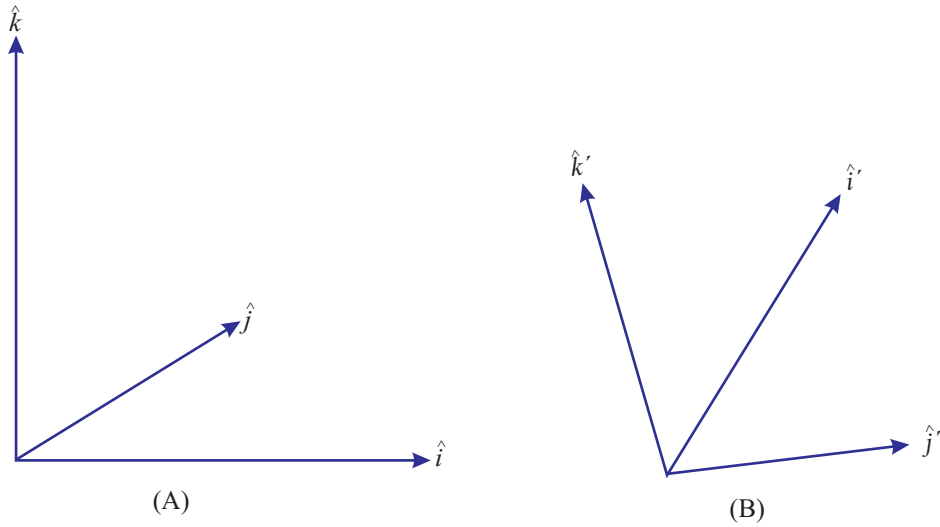


Figure 2-7: Orthonormal triad of vectors, (A) right handed triad, and (B) left handed triad.

Evidently, more than one (indeed, an *infinite* number of) orthonormal triad of vectors can be constructed in space. Fig. 2-7(B) depicts another such orthonormal triad, made up of unit vectors \hat{i}' , \hat{j}' , and \hat{k}' .

Looking carefully at fig. 2-7(A) and (B), one discerns a difference between the triads $\hat{i}-\hat{j}-\hat{k}$ and $\hat{i}'-\hat{j}'-\hat{k}'$. The *ordering* (i.e., the choice of which one of the three is taken *first*, which *second*, and which *third*) of the unit vectors is important here.

If one extends one's thumb, forefinger, and middle finger of the right hand in such a way that the thumb points in the direction of \hat{i} and the forefinger in the direction of \hat{j} in the above figure, then the middle finger will be found pointing in the direction of \hat{k} . On the other hand, if the thumb is made to point along \hat{i}' , and the forefinger along \hat{j}' , then the middle finger will be found to point in a direction *opposite* to \hat{k}' .

One then says that $\hat{i}-\hat{j}-\hat{k}$ and $\hat{i}'-\hat{j}'-\hat{k}'$ constitute a *right handed* and a *left handed* triad respectively.

2.5 Cartesian components of a vector

In fig. 2-8 OX, OY, OZ are three directed straight lines perpendicular to one another and $\hat{i}, \hat{j}, \hat{k}$ are unit vectors along these, forming a right handed orthonormal triad. One then says that OX, OY, OZ form a right handed *Cartesian co-ordinate system*. The *origin* of this system is at the point O.

Consider the vector **A**, represented by the directed line segment OA, in the context of this co-ordinate system. Perpendiculars, dropped from A on the three axes OX, OY, OZ, meet these axes at N_1, N_2, N_3 respectively. Let the directed distances of these points from the origin O be A_1, A_2, A_3 respectively.

In fig. 2-8, all the three directed distances are positive. On the other hand, if, say N_1 were located on the other side of the origin, then A_1 would have been negative. In other words, distances with appropriate signs are what I have termed directed distances.

In this instance, A_1, A_2, A_3 are the three Cartesian co-ordinates of the point A. At the same time, these are termed the *Cartesian components* of the vector **A** with reference to

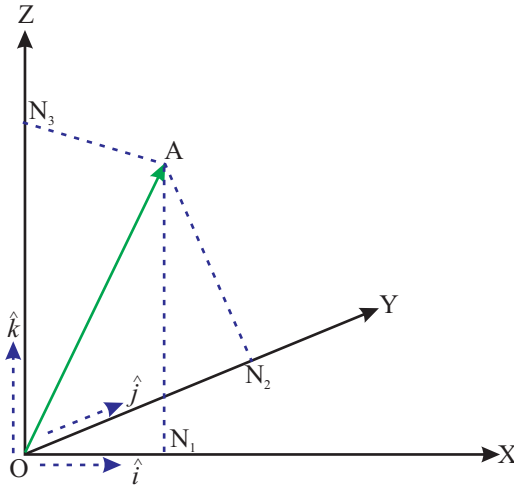


Figure 2-8: Illustrating the concept of Cartesian components of a vector; the vector \mathbf{A} is represented by the directed line segment OA , with the origin O (of a right handed Cartesian co-ordinate system) as the initial point; N_1, N_2, N_3 are feet of the perpendiculars dropped from A on OX, OY, OZ respectively.

the co-ordinate system $OXYZ$. One can express the vector \mathbf{A} in terms of these components and of the unit vectors $\hat{i}, \hat{j}, \hat{k}$ as

$$\mathbf{A} = A_1\hat{i} + A_2\hat{j} + A_3\hat{k}, \quad (2-18)$$

(check this out, making use of the definition of vector addition, and of multiplication with scalars.)

Problem 2-3

Show that the components of a vector (the Cartesian components are being referred to here as *components* in brief) with respect to a given co-ordinate system are *unique*, i.e., there cannot be two distinct sets of components.

Answer to Problem 2-3

HINT: If A'_1, A'_2, A'_3 be any other possible set of co-ordinates with reference to the same co-ordinate system, then

$$(A_1 - A'_1)\hat{i} + (A_2 - A'_2)\hat{j} + (A_3 - A'_3)\hat{k} = 0. \quad (2-19)$$

Taking scalar product with \hat{i} , one arrives at $A_1 = A'_1$, etc.

This means that, given a co-ordinate system, a vector is completely specified in terms of its Cartesian components in any given co-ordinate system, i.e., the ordered triplet of real numbers, (A_1, A_2, A_3) identifies the vector unambiguously.

The *same* vector is specified by a *different* ordered triplet of components in a different co-ordinate system. These may be, for instance, (A'_1, A'_2, A'_3) in a co-ordinate system determined by the triad $\hat{i}'\text{-}\hat{j}'\text{-}\hat{k}'$. However, there must exist some definite relation between the components (A_1, A_2, A_3) and (A'_1, A'_2, A'_3) because they correspond to the same vector in two different co-ordinate systems. Such a set of relations is termed a *vector transformation relation* (refer to sec. 2.9).

Numerous co-ordinate systems can be constructed using any given orthonormal triad $\hat{i}\text{-}\hat{j}\text{-}\hat{k}$, all of these being related to one another by parallel translation. All these co-ordinate systems lead to the *same* set of components for any given vector **A**. In other words, the Cartesian components of a vector are determined uniquely by the orthonormal triad under consideration.

Problem 2-4

Given two vectors $\mathbf{A} = \hat{i} - \hat{j} + 2\hat{k}$ and $\mathbf{B} = 3\hat{i} + \hat{j} - \hat{k}$, find a vector **C** such that $\mathbf{A} + 2\mathbf{B} - 3\mathbf{C} = 0$.

Answer to Problem 2-4

$$\mathbf{C} = \frac{7}{3}\hat{i} + \frac{1}{3}\hat{j}.$$

2.5.1 Two dimensional vectors

Think of the set of vectors represented by directed line segments that can, by parallel translation, all be made to lie in any given plane. This is termed the set of vectors lying

in that plane. Consider a pair of orthogonal unit vectors, say \hat{i}, \hat{j} in the plane (fig. 2-9). Then any vector belonging to the set under consideration can be expressed as

$$\mathbf{A} = A_1\hat{i} + A_2\hat{j}, \quad (2-20)$$

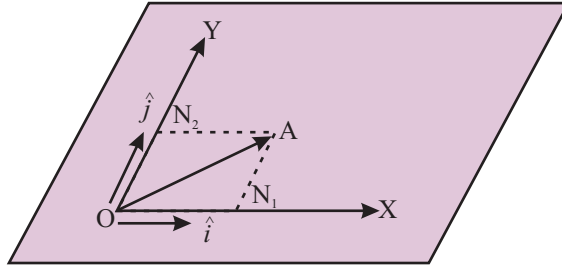


Figure 2-9: Cartesian components of a two dimensional vector; OX and OY are Cartesian axes in a plane along which the unit vectors are \hat{i}, \hat{j} ; the directed distances ON_1, ON_2 are the Cartesian components of the vector represented by the directed line segment OA lying in a given plane.

where the signed scalars A_1, A_2 are the Cartesian components of the vector with respect to a two dimensional co-ordinate system defined by \hat{i}, \hat{j} . These components are determined in a manner similar to the determination of the components of a three dimensional vector. Since any vector belonging to the set under consideration can be completely specified in terms of two components, this set forms a linear vector space of dimension two.

One can, in a similar manner, consider the set of vectors parallel to a line, where each of the vectors is specified by a single signed scalar and is a member of a vector space of dimension one.

2.5.2 Vector operations in terms of Cartesian components

Consider a right handed Cartesian co-ordinate system defined in terms of the orthonormal triad $\hat{i}-\hat{j}-\hat{k}$. Let the components of a vector \mathbf{A} in this co-ordinate system (or, more precisely, with respect to the above triad) be A_1, A_2, A_3 . If p be any scalar then the components of the vector $p\mathbf{A}$, obtained by multiplying the vector \mathbf{A} with the scalar p ,

will be pA_1, pA_2, pA_3 :

$$p\mathbf{A} = pA_1\hat{i} + pA_2\hat{j} + pA_3\hat{k}. \quad (2-21)$$

In other words, the components of $p\mathbf{A}$ are simply p times the components of \mathbf{A} .

Again, if \mathbf{B} be any vector with components B_1, B_2, B_3 with respect to the same orthonormal triad then the components of the vector $\mathbf{A} + \mathbf{B}$ will be $A_1 + B_1, A_2 + B_2, A_3 + B_3$, i.e.,

$$\mathbf{A} + \mathbf{B} = (A_1 + B_1)\hat{i} + (A_2 + B_2)\hat{j} + (A_3 + B_3)\hat{k}. \quad (2-22)$$

This implies that the components of the vector sum of a number of vectors are the sums of the corresponding components of those vectors.

(Check these results out.)

2.5.3 Scalar product of two vectors in terms of Cartesian components

Let A_1, A_2, A_3 and B_1, B_2, B_3 be the components of any two vectors \mathbf{A} and \mathbf{B} with respect to the orthonormal triad $\hat{i}-\hat{j}-\hat{k}$. Then, making use of features, mentioned above, of (a) multiplication of a vector with a scalar, (b) vector addition, and (c) scalar product of two vectors, one finds

$$\mathbf{A} \cdot \mathbf{B} = A_1B_1 + A_2B_2 + A_3B_3 = \sum_{i=1}^3 A_iB_i. \quad (2-23)$$

This is a convenient formula giving the scalar product in terms of its Cartesian components.

Problem 2-5

Establish formula (2-23).

Answer to Problem 2-5

HINT: $\mathbf{A} \cdot \mathbf{B} = (A_1\hat{i} + A_2\hat{j} + A_3\hat{k}) \cdot (B_1\hat{i} + B_2\hat{j} + B_3\hat{k})$. Now use $\hat{i} \cdot \hat{i} = 1$, $\hat{i} \cdot \hat{j} = 0$ etc., and the features of scalar product of vectors mentioned in sec. 2.4.1.

2.5.4 Direction cosines relating to a unit vector

Consider a Cartesian co-ordinate system determined by the orthonormal triad $\hat{i}-\hat{j}-\hat{k}$. Now think of a unit vector \hat{n} , represented by the directed line segment OP with the initial point at the origin O (recall that the initial point can be taken anywhere by making use of parallel translation). Let the *direction cosines* of the directed straight line along OP be l_1, l_2, l_3 . This means that, if $\theta_1, \theta_2, \theta_3$ be the angles made with OX, OY, and OZ respectively, by the directed line along OP (fig. 2-10), then $l_1 = \cos \theta_1$, $l_2 = \cos \theta_2$, $l_3 = \cos \theta_3$. Then, the components of \hat{n} with respect to the triad $\hat{i}-\hat{j}-\hat{k}$ will be l_1, l_2, l_3 , i.e.,

$$\hat{n} = l_1\hat{i} + l_2\hat{j} + l_3\hat{k}, \quad (2-24)$$

(check this out). Using the fact that \hat{n} is a unit vector, one observes that the direction cosines (l_1, l_2, l_3) of a directed line satisfy

$$l_1^2 + l_2^2 + l_3^2 = 1, \quad (2-25)$$

a known result in geometry.

Problem 2-6

A directed line makes angles $\frac{\pi}{3}, \frac{\pi}{3}$ with the x- and y-axes of a co-ordinate system, and points into the octant defined by the positive directions of the x- and y-axes, and the negative direction of the z-axis; find the unit vector along the directed line.

Answer to Problem 2-6

HINT: The direction cosines of the line are $l_1 = \frac{1}{2}$, $l_2 = \frac{1}{2}$, l_3 (say), where l_3 is of negative sign;

now make use of the fact that these are the components of the required unit vector, which implies

$$l_3 = -\frac{1}{\sqrt{2}}.$$

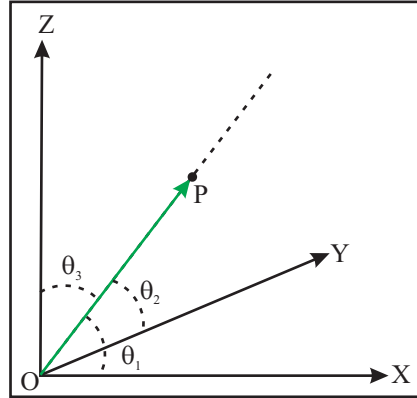


Figure 2-10: Illustrating the concept of direction cosines; The directed line segment OP represents a unit vector, making angles $\theta_1, \theta_2, \theta_3$ with the co-ordinate axes OX, OY, OZ respectively; any directed line along OP is then said to be characterized by direction cosines $l_1 = \cos \theta_1, l_2 = \cos \theta_2, l_3 = \cos \theta_3$.

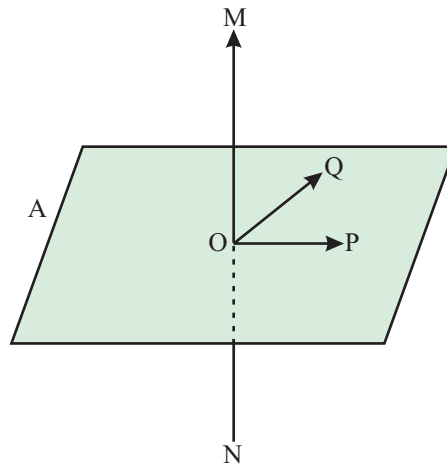


Figure 2-11: Illustrating the concept of cross product of two vectors; the cross product of \vec{OP} and \vec{OQ} is a vector along OM .

2.6 The vector product of two vectors

In fig. 2-11 the directed line segments OP and OQ represent two vectors, say, \mathbf{A} and \mathbf{B} , and the plane containing these two segments is denoted by \mathcal{A} . The normal to this plane through the point O is OM . Now imagine a right handed cork-screw being rotated, the sense of rotation being *from* OP *to* OQ . Then, in the present instance, the cork-screw will advance in the direction of the directed line along OM . One then says that the direction of OM is related to the directions of OP and OQ *in the right handed sense* (see sec. 2.8). Choosing the direction related to the directions of \mathbf{A} and \mathbf{B} in the right handed sense in this manner, imagine a unit vector \hat{n} along this direction. Let the angle between \mathbf{A} and \mathbf{B} be θ .

Recall that two directed lines generate, between them, *two* angles. If one of these be named θ , then the other would be $2\pi - \theta$. In the present context, the *smaller* of the two angles has been taken as θ . While talking of the scalar product of two vectors, I mentioned that *either* of the two angles could be taken as θ . By contrast, one has to take the smaller angle in defining the vector product.

Now think of a vector whose magnitude is $|\mathbf{A}||\mathbf{B}|\sin\theta$ and direction is along OM . This vector, which can be written as $(|\mathbf{A}||\mathbf{B}|\sin\theta)\hat{n}$, is defined as the *vector product* of the vectors \mathbf{A} and \mathbf{B} (the name *cross product* is also used). One writes

$$\mathbf{A} \times \mathbf{B} = (|\mathbf{A}||\mathbf{B}|\sin\theta)\hat{n}. \quad (2-26)$$

As an example, think of a right handed orthonormal triad of vectors, \hat{i} , \hat{j} , \hat{k} . Then, following the above rule of formation of the vector product of two vectors, one obtains

$$\begin{aligned} \hat{i} \times \hat{j} &= \hat{k}, \\ \hat{j} \times \hat{k} &= \hat{i}, \\ \hat{k} \times \hat{i} &= \hat{j}, \end{aligned} \quad (2-27)$$

(check this out.)

2.6.1 Features of the vector product

Let \mathbf{A} , \mathbf{B} , \mathbf{C} be any three vectors and p , q any two scalars. Then, a few features of the cross product as defined above may be stated in terms of these as follows.

1.

$$\mathbf{A} \times \mathbf{B} = -\mathbf{B} \times \mathbf{A}. \quad (2-28)$$

In other words, the *order* of the two vectors is important in the vector product, which is *antisymmetric* in the two vectors. This is to be compared with the scalar product of the two vectors where the order can be interchanged without altering the value of the product.

2.

$$\mathbf{A} \times (p\mathbf{B} + q\mathbf{C}) = p\mathbf{A} \times \mathbf{B} + q\mathbf{A} \times \mathbf{C}, \quad (2-29)$$

i.e., in other words, the vector product is *distributive* with respect to vector addition and is, more precisely, *linear* in either of the two vectors featuring in the product.

3. If the directed line segments representing, say, \mathbf{A} and \mathbf{B} are either parallel or anti-parallel to each other, i.e., \mathbf{B} is of the form $p\mathbf{A}$ with p either a positive (parallel) or a negative (anti-parallel) scalar, then

$$\mathbf{A} \times \mathbf{B} = 0. \quad (2-30)$$

4. If the unit vectors \hat{i} , \hat{j} , \hat{k} form a right handed orthonormal triad, and if the components of \mathbf{A} , \mathbf{B} with respect to this triad be respectively (A_1, A_2, A_3) , and (B_1, B_2, B_3) , then

$$\mathbf{A} \times \mathbf{B} = (A_2B_3 - A_3B_2)\hat{i} + (A_3B_1 - A_1B_3)\hat{j} + (A_1B_2 - A_2B_1)\hat{k}. \quad (2-31)$$

In other words, the cross product of two vectors can be expressed in terms of their

Cartesian components relative to a given right handed orthonormal triad, along with the unit vectors of that triad.

Problem 2-7

Establish the result (2-31).

Answer to Problem 2-7

HINT: by making use of the above features of the cross-product, and (2-27).

Digression: Axial and polar vectors. From the mathematical point of view, the vector product of two vectors, say \mathbf{A} and \mathbf{B} , does not, strictly speaking, belong to the same linear vector space as the two vectors themselves. Rather, the vector product $\mathbf{A} \times \mathbf{B}$ is properly described as an *antisymmetric tensor of rank two*. Another way to refer to the vector product is to call to it an *axial vector*. However, in this book we shall continue to refer to the vector product as a vector since the mathematical distinction involved will not, in most cases, be of direct relevance for.

Examples of axial vectors in physics are angular momentum, torque, and magnetic intensity. In order to distinguish these from other commonly occurring vectors like velocity, force, momentum, and electric intensity, the latter are referred to as *polar vectors*. Axial vectors are seen to appear in mathematical descriptions as vector products of polar vectors.

Thus, if \mathbf{A} and \mathbf{B} are polar vectors, then $\mathbf{A} \times \mathbf{B}$ will be an axial vector.

Problem 2-8

Consider the vector $\mathbf{A} = \hat{i} - \hat{j}$, and a second vector \mathbf{B} such that $\mathbf{A} \times \mathbf{B} = 2\hat{k}$ and $\mathbf{A} \cdot \mathbf{B} = 0$. Find \mathbf{B} .

Answer to Problem 2-8

HINT: Since $\mathbf{A} \times \mathbf{B}$ points along the z-axis, \mathbf{B} has to lie in the x-y plane, as does \mathbf{A} . Writing \mathbf{B} in the form $\mathbf{B} = a\hat{i} + b\hat{j}$ and observing that \mathbf{B} is orthogonal to \mathbf{A} , one obtains $a = b$. Then, making

use of the vector product, one gets $\mathbf{B} = \hat{i} + \hat{j}$.

2.6.2 Vector expression for a planar area

Analogous to the fact that a straight line in three dimensional space is characterized by a certain direction, a plane is also characterized by a certain *orientation*. Fig. 2-12 depicts a triangle in two different orientations. One way to distinguish between the two orientations is to specify the directions of the *normals* drawn to the planes of the triangles.

Let the area of the triangle be A and the unit vector along the normal to the plane of the triangle in any given orientation be \hat{n} (fig. 2-12(A)). Then a convenient way to express the area *along with* the orientation of the plane of the triangle is to refer to the vector

$$\Delta = A\hat{n}. \quad (2-32)$$

This vector is termed the *vector area* of the triangle. Its norm specifies the area of the triangle while its direction gives the orientation of the plane of the triangle.

1. A triangle is necessarily contained in a plane. However, a polygon of more than three sides need not be contained in a plane. For a planar polygon, however, the concept of vectorial area can be introduced as in the case of a triangle.
2. There can be *two* unit vectors normal to a given plane. In fig. 2-12(A), for instance, the unit vector \hat{n} could point from O to M instead of pointing from O to N. The criterion for choosing one from among these two possible unit vectors is often related to the *sense* of traversal (clockwise or anti-clockwise) of some closed contour in the plane under consideration. Thus, in fig. 2-12(A) the curved arrow indicates an anti-clockwise traversal of the closed contour made of the sides of the triangle. The unit vector \hat{n} is then chosen to be related to this sense of traversal by the *right hand rule*: a right handed cork-screw made to rotate in the sense of the bent arrow advances in the direction of \hat{n} .

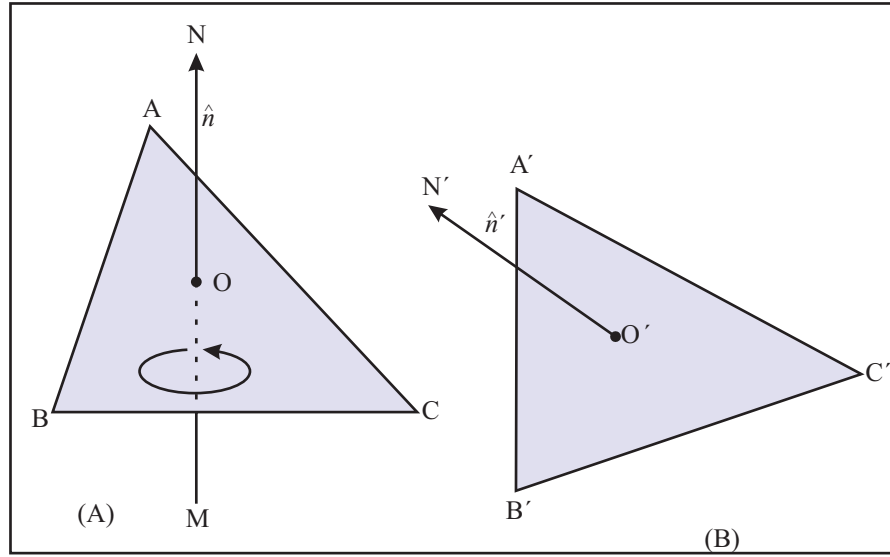


Figure 2-12: Specifying the orientation of a plane with the help of the normal drawn to the plane; two different orientations of a triangle are shown in (A) and (B), along with the unit normals.

Problem 2-9

If \mathbf{A} and \mathbf{B} be the vectors represented by the directed line segments AB and AC respectively, say, in fig. 2-12(A), then show that the vectorial area of the triangle is given by

$$\Delta = \frac{1}{2} \mathbf{A} \times \mathbf{B}. \quad (2-33)$$

Answer to Problem 2-9

HINT: The magnitude of the given expression is $\frac{1}{2} AB \cdot AC \sin \theta$, where θ is the angle between AB and AC, and the direction is perpendicular to the plane of the triangle.

2.7 Scalar and vector components of a vector

Fig. 2-13((A), (B)) depicts a vector \mathbf{A} represented by the directed line segment \vec{AB} , and a unit vector \hat{n} along any given line XX' . Let M and N be the feet of the perpendiculars dropped from A and B respectively on XX' . Then the directed distance MN is referred to as the scalar component (or simply the component) of \mathbf{A} along \hat{n} . This directed distance

is represented by a scalar with a sign. For instance, the scalar component of \mathbf{A} along \hat{n} is positive in fig. 2-13(A) and negative in fig. 2-13(B). The mathematical expression for the scalar component is given by

$$\text{scalar component of } \mathbf{A} \text{ along } \hat{n} = \mathbf{A} \cdot \hat{n} = A \cos \theta, \quad (2-34)$$

where θ is the angle between the oriented lines along \vec{AB} and \hat{n} . The above is also commonly referred to as the component of the vector \mathbf{A} along any directed line, like XX' in the figure, parallel to \hat{n} , or along any vector parallel to \hat{n} .

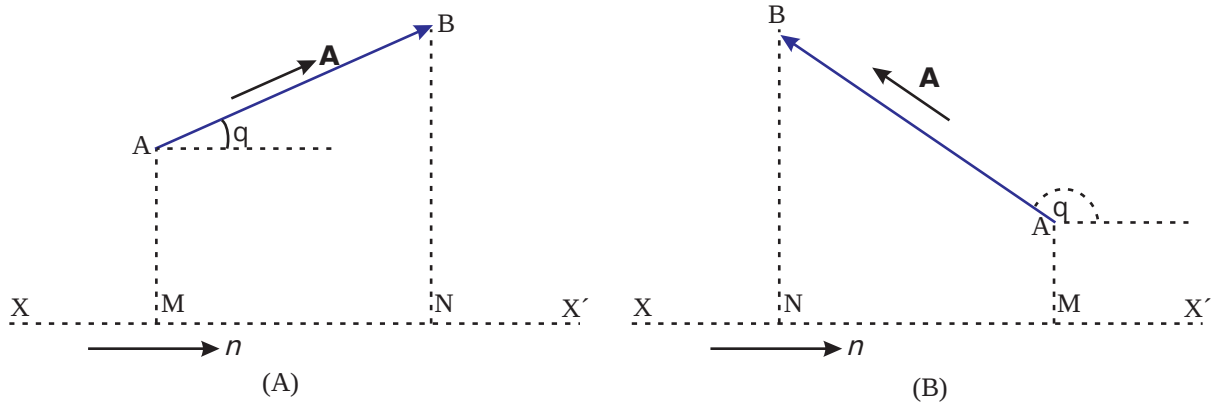


Figure 2-13: Illustrating the idea of the scalar component (or, simply, the component) of a vector \mathbf{A} (represented by the directed line segment \vec{AB}) along a given unit vector (\hat{n}); the component, given by the expression $A \cos \theta$ (eq. (2-34)), is represented by the directed distance MN or, equivalently, by a scalar with a sign, being positive in (A), and negative in (B).

Thus, for instance, the x-component of any given vector \mathbf{A} with reference to any Cartesian co-ordinate system with axes along unit vectors $\hat{i}, \hat{j}, \hat{k}$, is $A_x = \mathbf{A} \cdot \hat{i}$.

Consider now the vector $(\mathbf{A} \cdot \hat{n})\hat{n}$, which is represented by the directed line segment \vec{MN} in fig. 2-13(A), (B). This is termed the vector component of \mathbf{A} along \hat{n} (at times, just the term 'component' is used).

Considering two vectors \mathbf{A} and \mathbf{B} , the component of \mathbf{A} along \mathbf{B} stands for the (vector) component of \mathbf{A} along the unit vector $\hat{b}(= \frac{\mathbf{B}}{|\mathbf{B}|})$ in the direction of \mathbf{B} .

Fig. 2-14 illustrates the concept of the vector component of a vector (**A**) in a direction *perpendicular* to another vector (**B**) or to a directed line, as explained in the problem below. In this figure, the directed line segment OP represents the vector component of the vector **A** along the vector **B** (in the example shown, the scalar component of **A** along **B** is of negative sign). The directed line segment PA then represents the vector component of **A** perpendicular to **B**.

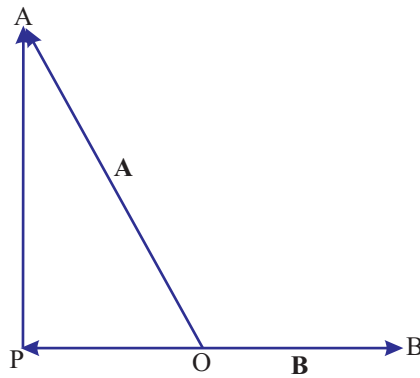


Figure 2-14: Illustrating the idea of the vector component of a vector **A**, represented by the directed line segment OA, along another vector **B**, represented by OB; OP represents the vector component of **A** along **B** which, in the present instance, is directed opposite to **B**; the directed line segment PA then represents the vector component of **A** perpendicular to **B**.

Problem 2-10

Looking at fig. 2-14, show that the vector component of **A** perpendicular to **B** is given by $(\mathbf{A} - \frac{\mathbf{A} \cdot \mathbf{B}}{B^2} \mathbf{B})$.

Answer to Problem 2-10

HINT: According to the definition of the vector component of a vector along another vector, the directed line segment OP stands for the vector $(\mathbf{A} \cdot \hat{b})\hat{b}$, where $\hat{b} = \frac{\mathbf{B}}{B}$ is the unit vector along **B**. The required result then follows from the fact that the vector sum of \vec{OP} and \vec{PA} equals the given vector **A**.

2.8 The right hand rule

The *right hand rule*, referred to in sections 2.4.2 and 2.6 will feature repeatedly in this book, especially in the context of mechanics and of the magnetic effect of currents. In sec. 2.4.2, the right hand rule was explained by referring to an imagined situation in which the thumb, the forefinger, and the middle finger of one's right hand are extended. Considering now three directed lines, say, L_1 , L_2 , L_3 respectively, pointing along the three extended fingers, L_3 is said to be related to L_1 , and L_2 by the right hand rule (see fig. 2-15(A)) or, in the right handed sense.

The rule does not actually need that all the three directed lines make an angle of $\frac{\pi}{2}$ with each other. For instance, the directed line L_3 in fig. 2-15(B) is related to L_1 and L_2 in the right handed sense, where the angle between L_1 and L_2 differs from $\frac{\pi}{2}$, and where this can be explained by referring to the way the rule was stated in sec. 2.6: if a right handed cork-screw be imagined to be rotated in the sense *from* L_1 *to* L_2 (through the smaller of the two angles made between the two lines), it advances along the direction of L_3 .

One can describe the situation indicated in fig. 2-15(A), (B) in an alternative way by saying that the direction of L_3 and the sense of rotation (from L_1 to L_2 ; indicated by bent arrows in the figure) are related to each other by the right hand rule.

Thus, the right hand rule describes a relation between the directions of three directed lines or three vectors, as also a relation between a given direction in space and a given sense of rotation. It may also describe a relation (see fig. 2-15(C)) between a given point P , a directed line L_1 , and a second directed line L_2 passing through P . Imagine a perpendicular being dropped from P on to L_1 , with N (not shown in the figure) as the foot of the perpendicular. If, then, L_2 is related to the directed lines PN and L_1 (now imagined to be parallel transported to pass through P) in the right handed sense, one can state that L_2 is related to the point P and to the directed line L_1 (in its original position) in the right handed sense. Note that what is actually involved here is again the relation between a given sense of rotation (the sense of rotation around P described by L_1) and a direction in space (that of L_2).

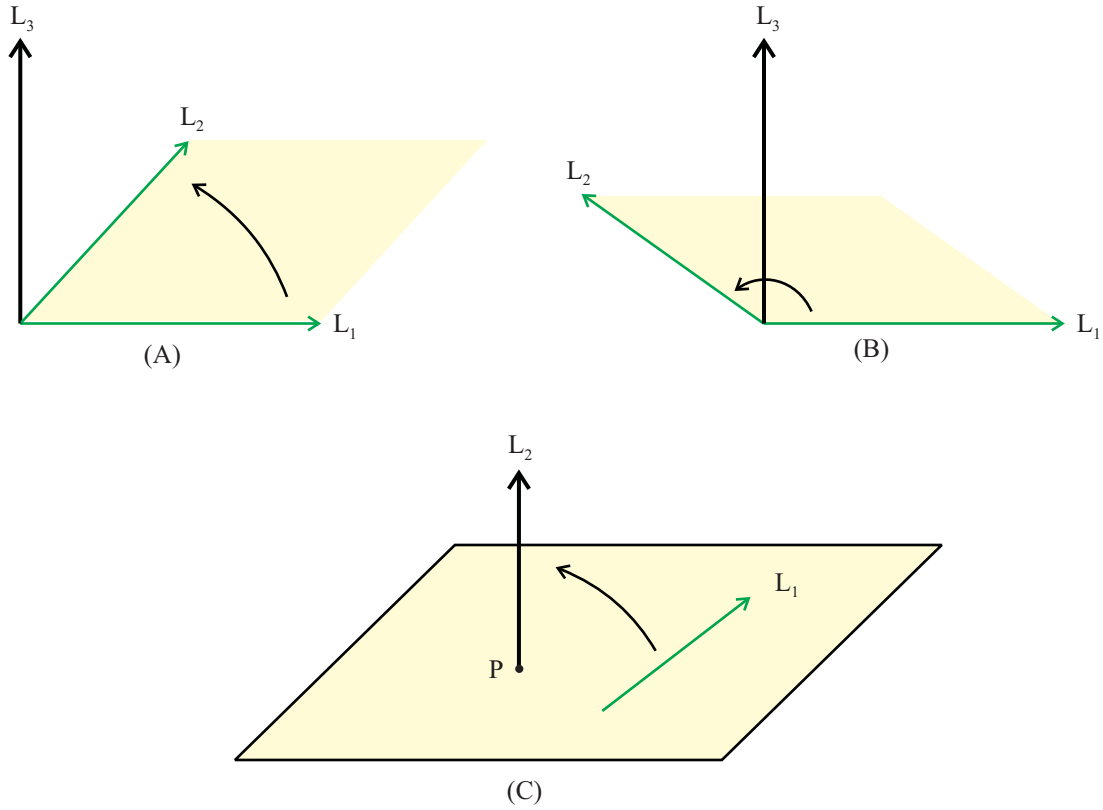


Figure 2-15: Illustrating the right hand rule; (A), (B) the direction of the directed line L_3 is related to the directions of the lines L_1 and L_2 by the right hand rule; in (A), the angle between L_1 and L_2 is $\frac{\pi}{2}$ while in (B), the angle differs from $\frac{\pi}{2}$; (C) the direction of L_2 is related to the point P and the direction of L_1 by the right hand rule; in each case the right hand rule actually expresses a relation between a given direction in space and a sense of rotation; the latter is indicated by bent arrows in (A), (B), and (C).

2.9 Transformation of vectors

Imagine a Cartesian co-ordinate system with axes OX, OY, OZ parallel to the unit vectors \hat{e}_1 , \hat{e}_2 , \hat{e}_3 , and a vector \mathbf{A} with components a_1, a_2, a_3 along these:

$$\mathbf{A} = a_1 \hat{e}_1 + a_2 \hat{e}_2 + a_3 \hat{e}_3 = \sum_{i=1}^3 a_i \hat{e}_i. \quad (2-35)$$

The notation here is different from the one we have been using so long (and will use later as well) in this chapter where the triad of unit vectors is denoted by $\hat{i}, \hat{j}, \hat{k}$. The present notation allows one to write compact formulae by making use of the summa-

tion symbol.

Imagine now a second set of Cartesian axes OX', OY', OZ' , parallel to the unit vectors $\hat{e}'_1, \hat{e}'_2, \hat{e}'_3$, as in fig. 2-16, with respect to which the same vector \mathbf{A} has components a'_1, a'_2, a'_3 :

$$\mathbf{A} = a'_1 \hat{e}'_1 + a'_2 \hat{e}'_2 + a'_3 \hat{e}'_3 = \sum_{i=1}^3 a'_i \hat{e}'_i. \quad (2-36)$$

In the figure, we have taken the origins of the two sets of axes to be coincident for the sake of convenience, since the description of a vector does not depend on the choice of the origin.

The system $OX'Y'Z'$ can be obtained from the system $OXYZ$ by a rotation about some axis passing through O and, possibly, an inversion of one or all three of the axes about O , where the inversion is necessary if one of the two Cartesian systems is a right-handed one while the other is left-handed (i.e., the two systems differ in the ‘handed-ness’).

Making use of the formulae (2-35), (2-36), one can work out the coefficients a'_1, a'_2, a'_3 in terms of a_1, a_2, a_3 . For instance, taking the scalar product of both sides of (2-35) and also of the two sides of (2-36) with \hat{e}'_1 , one obtains

$$a'_1 = \hat{e}'_1 \cdot \hat{e}_1 a_1 + \hat{e}'_1 \cdot \hat{e}_2 a_2 + \hat{e}'_1 \cdot \hat{e}_3 a_3, \quad (2-37a)$$

where the left hand side is obtained by observing that \hat{e}'_1 is orthogonal to both of \hat{e}'_2 and \hat{e}'_3 . Similar expressions for a'_2 and a'_3 can be obtained by taking scalar products with \hat{e}'_2, \hat{e}'_3 respectively. All the three expressions so obtained can be written down in the following compact form using the summation symbol,

$$a'_i = \sum_{j=1}^3 \hat{e}'_i \cdot \hat{e}_j a_j \quad (i = 1, 2, 3), \quad (2-37b)$$

(check this out). This gives the transformation from the set of components $\{a_i\}$ of the vector \mathbf{A} to the transformed set $\{a'_i\}$. Defining the *transformation coefficients*

$$t_{ij} = \hat{e}'_i \cdot \hat{e}_j \quad (i, j = 1, 2, 3), \quad (2-37c)$$

one arrives at the still more compact form describing the transformation,

$$a'_i = \sum_j t_{ij} a_j. \quad (2-37d)$$

The transformation coefficients t_{ij} , which depend on the orientation of the axes of the system $OX'Y'Z'$ relative to $OXYZ$ (and hence are the same for all vectors), form the elements of a 3×3 *orthogonal matrix* for which the following relations hold

$$\sum_k t_{ik} t_{jk} = \delta_{ij}, \quad \sum_k t_{ki} t_{kj} = \delta_{ij}, \quad (2-38a)$$

where δ_{ij} stands for the *Krönecker delta* symbol, defined as

$$\delta_{ij} = 0 \text{ (for } i \neq j), \quad \delta_{ij} = 1 \text{ (for } i = j) \text{ (} i, j = 1, 2, 3). \quad (2-38b)$$

(Check this out; for the first relation in (2-38a), consider the scalar product $\hat{e}'_i \cdot \hat{e}'_j$ and make use of the relation $\hat{e}'_i = \sum_j t_{ij} \hat{e}_j$ ($j = 1, 2, 3$); the second relation is derived by interchanging the primed and unprimed objects; the matrix T with elements $\{t_{ij}\}$ is commonly referred to as the *transformation matrix* in the context under consideration; if you are not conversant with matrices, skip ahead of problem 2-11 below; but I suggest you acquire a working knowledge in matrices (not difficult), and come back to it; we will be needing matrices especially in chapter 18).

Analogous to the transformation equations (2-37b), (2-37d), one can derive the inverse transformation from the set of components $\{a'_i\}$ to the set $\{a_i\}$ as

$$a_i = \sum_j t_{ji} a'_j \text{ (} i = 1, 2, 3). \quad (2-39)$$

Problem 2-11

Consider a right handed co-ordinate system S made up of axes OX, OY, OZ , as in fig. 2-17, and a second system S' with axes OX', OY', OZ' , obtained by a rotation about OZ by an angle ϕ (thus,

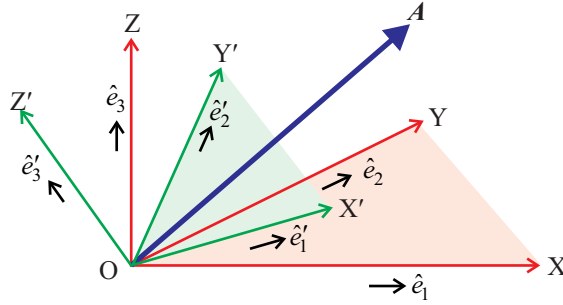


Figure 2-16: Illustrating the idea of transformation of the components of a vector under a change of Cartesian axes; $OXYZ$ and $OX'Y'Z'$ are two sets of Cartesian axes, where both are shown to be right handed for the sake of illustration; \hat{e}_i ($i = 1, 2, 3$) are unit vectors along the axes of the first system while \hat{e}'_i ($i = 1, 2, 3$) are unit vectors for the second system; the vector \mathbf{A} has components a_i with respect to the first set, and a'_i with respect to the second set ($i = 1, 2, 3$); the two sets of components are related as in (2-37b), (2-37d); the inverse transformation from the set $\{a'_i\}$ to the set $\{a_i\}$ is of the same type and is given by eq. (2-39).

OZ' coincides with OZ), as shown in the figure. Obtain the transformation matrix expressing the components of a vector in S' in terms of those in S .

Answer to Problem 2-11

Consider any point P with position vector \mathbf{r} , whose components in the two systems are (x, y, z) and (x', y', z') . From the geometry of the figure, one obtains

$$x' = x \cos \phi + y \sin \phi, \quad y' = -x \sin \phi + y \cos \phi, \quad z' = z, \quad (2-40)$$

(check this out). This gives the transformation matrix (the components of all vectors transform in a similar manner)

$$T = \begin{pmatrix} \cos \phi & \sin \phi & 0 \\ -\sin \phi & \cos \phi & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (2-41)$$

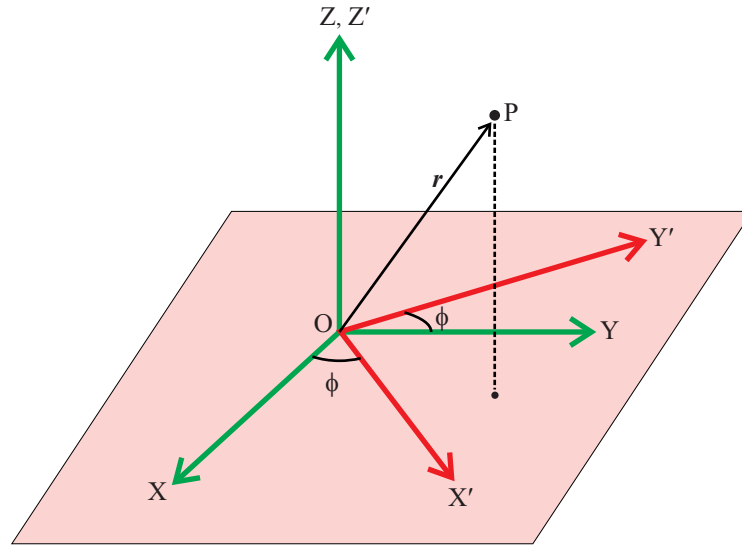


Figure 2-17: Illustrating the co-ordinate systems S and S' mentioned in problem 2-11; S' is obtained from S by a rotation through an angle ϕ about OZ (thus, OZ' coincides with OZ ; all rotations are measured in the right handed sense, see sec. 2.8); P is a point with position vector \mathbf{r} having components (x, y, z) in S and (x', y', z') in S' . These are related by the transformation matrix T of eq. (2-41).

2.10 Scalar and vector triple products

2.10.1 Scalar triple product

Fig. 2-18 depicts a parallelepiped where the directed line segments \vec{OA} , \vec{OB} , \vec{OC} are along three adjacent sides of the parallelepiped. Let the vectors represented by these directed line segments be denoted by \mathbf{A} , \mathbf{B} , \mathbf{C} respectively. In the figure, α denotes the angle between $\mathbf{B} = \vec{OB}$ and $\mathbf{C} = \vec{OC}$, and β the angle between $\mathbf{A} = \vec{OA}$ and \hat{n} , where \hat{n} stands for the unit vector along $\mathbf{B} \times \mathbf{C}$.

Now consider the scalar product of \mathbf{A} with the vector product of \mathbf{B} and \mathbf{C} . One obtains for this quantity the expression

$$\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}) = |\mathbf{A}| |\mathbf{B}| |\mathbf{C}| \sin \alpha \cos \beta. \quad (2-42)$$

(Check the above statement out.)

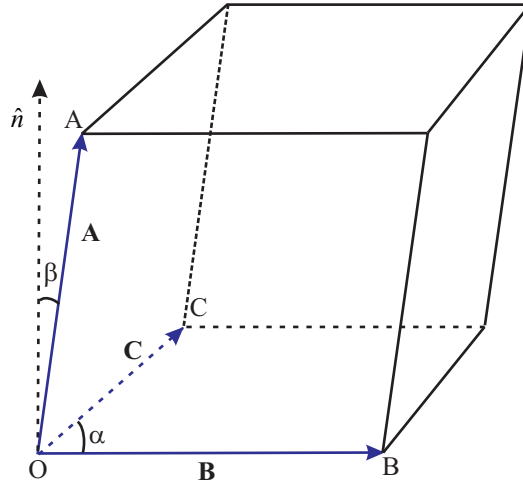


Figure 2-18: Illustrating the idea of the scalar triple product; the directed line segments OA, OB, OC representing vectors \mathbf{A} , \mathbf{B} , \mathbf{C} form three adjacent sides (OA, OB, OC) of a parallelepiped; the scalar triple product $\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C})$ then corresponds to the oriented volume of the parallelepiped; \hat{n} is the unit vector along $\mathbf{B} \times \mathbf{C}$.

The expression on the right hand side of the above equation is referred to as the *oriented volume* of the parallelepiped under consideration because it is a scalar with a sign (either positive or negative) whose magnitude equals the volume of the parallelepiped (check this out) while the sign depends on the relative orientations of the vectors \mathbf{A} , \mathbf{B} , and \mathbf{C} .

For any three given vectors \mathbf{A} , \mathbf{B} , \mathbf{C} , the product $\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C})$, formed according to the rules of formation of vector and scalar products of vectors, can thus be interpreted as the oriented volume of a parallelepiped with three adjacent sides corresponding to the vectors as explained above, and is referred to as the *scalar triple product* of the vectors, taken in that order. Considering a right handed Cartesian co-ordinate system in which the vectors are represented by components A_i, B_i, C_i ($i = 1, 2, 3$), the scalar triple product is given by

$$\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}) = A_1(B_2C_3 - B_3C_2) + A_2(B_3C_1 - B_1C_3) + A_3(B_1C_2 - B_2C_1). \quad (2-43)$$

(Check eq. (2-43) out.)

Problem 2-12

Given three vectors \mathbf{A} , \mathbf{B} , \mathbf{C} , show that

$$\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}) = \mathbf{B} \cdot (\mathbf{C} \times \mathbf{A}) = -\mathbf{C} \cdot (\mathbf{B} \times \mathbf{A}). \quad (2-44)$$

.

Answer to Problem 2-12

HINT: Compare each of these expressions with the right hand side of eq. (2-43). The triple product $\mathbf{C} \cdot (\mathbf{B} \times \mathbf{A})$ gives the oriented volume of the parallelepiped referred to above, with a sign opposite to what the first two expressions in the above equation correspond to.

2.10.2 Vector triple product

Given any three vectors \mathbf{A} , \mathbf{B} , \mathbf{C} , consider first the vector product of \mathbf{B} with \mathbf{C} , and then the vector product of \mathbf{A} with the resulting vector. The result of this operation is a vector, given by the expression $\mathbf{A} \times (\mathbf{B} \times \mathbf{C})$, and is termed a *vector triple product* formed of the three given vectors. Making use of the expression of the vector product in terms of Cartesian components, one can express the vector triple product in terms of the components A_i, B_i, C_i ($i = 1, 2, 3$) of the three vectors with reference to a right handed Cartesian co-ordinate system, which works out to

$$\begin{aligned} \mathbf{A} \times (\mathbf{B} \times \mathbf{C}) &= (A_2(B_1C_2 - B_2C_1) - A_3(B_3C_1 - B_1C_3))\hat{i} \\ &\quad + (A_3(B_2C_3 - B_3C_2) - A_1(B_1C_2 - B_2C_1))\hat{j} \\ &\quad + (A_1(B_3C_1 - B_1C_3) - A_2(B_2C_3 - B_3C_2))\hat{i}. \end{aligned} \quad (2-45)$$

(Check this out).

The vector triple product satisfies the following relation

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = (\mathbf{A} \cdot \mathbf{C})\mathbf{B} - (\mathbf{A} \cdot \mathbf{B})\mathbf{C}. \quad (2-46)$$

Problem 2-13

Check eq. (2-46) out.

Answer to Problem 2-13

(a) Evaluate both sides in any right handed Cartesian co-ordinate system.

(b) One can also establish the identity without reference to a co-ordinate system. Note that $\mathbf{A} \times (\mathbf{B} \times \mathbf{C})$ has to lie in the plane defined by \mathbf{B} and \mathbf{C} and hence has to be of the form $\alpha\mathbf{B} + \beta\mathbf{C}$, where α and β are scalars (reason this out; a vector of the form $\alpha\mathbf{B} + \beta\mathbf{C}$ is referred to as a *linear combination* of the vectors \mathbf{B} and \mathbf{C}). Now take the scalar product with \mathbf{A} to show that α and β have to be of the form $\lambda(\mathbf{A} \cdot \mathbf{C})$ and $-\lambda(\mathbf{A} \cdot \mathbf{B})$, where λ is independent of \mathbf{A} , \mathbf{B} , \mathbf{C} . Finally, choose \mathbf{B} and \mathbf{C} to be perpendicular to each other and each to be unit length, and take $\mathbf{A} = \mathbf{B}$ (say), to obtain $\lambda = 1$.

Problem 2-14

Consider a vector $\mathbf{A} = 2\hat{i} - \hat{j} + 2\hat{k}$ and a second vector \mathbf{B} perpendicular to \mathbf{A} . If $\mathbf{A} \times \mathbf{B} = -5\hat{i} - 2\hat{j} + 4\hat{k}$, find \mathbf{B} in terms of \hat{i} , \hat{j} , \hat{k} .

Answer to Problem 2-14

HINT: Making use of eq. (2-46) and of the fact that $\mathbf{A} \cdot \mathbf{B} = 0$, show that $A^2\mathbf{B} = -\mathbf{A} \times (\mathbf{A} \times \mathbf{B})$.

Hence obtain $\mathbf{B} = 2\hat{j} + \hat{k}$.

2.11 Vector function of a scalar variable

Consider a scalar variable, say, x , that can take up real values in any given interval, say $a < x < b$, and assume that, for each specified value of x in this interval, one has a vector \mathbf{A} uniquely defined for that x . One then has a *set* of vectors, where a typical member of the set, corresponding to a given value of x can be denoted by the symbol $\mathbf{A}(x)$. One thereby has a vector *function* of the scalar variable x , defined by the said unique correspondence between x and $\mathbf{A}(x)$.

Referring to any Cartesian co-ordinate system, a vector function $\mathbf{A}(x)$ can be specified in terms of three scalar functions specifying the components of $\mathbf{A}(x)$. For instance, with the scalar variable x defined over the interval $-\infty < x < \infty$, the expression

$$\mathbf{A}(x) = \sin x \hat{i} + \cos x \hat{j} + e^{-x^2} \hat{k}, \quad (2-47)$$

defines a vector function of x since, for each chosen value of x it gives us a uniquely defined vector $\mathbf{A}(x)$. More generally, a vector function $\mathbf{A}(x)$ defines three scalar functions $A_i(x)$ ($i = 1, 2, 3$), and *vice versa*, by means of the relation

$$\mathbf{A}(x) = A_1(x)\hat{i} + A_2(x)\hat{j} + A_3(x)\hat{k}. \quad (2-48)$$

Given a number of vector functions $\mathbf{A}_i(x)$ ($i = 1, 2, \dots, N$), all defined over some common interval of the scalar variable x , one can define a vector function $\mathbf{A}(x)$ as the sum of all these functions, defined over that common interval, as

$$\mathbf{A}(x) = \sum_{i=1}^N \mathbf{A}_i(x). \quad (2-49)$$

The product of a vector function \mathbf{A} of the scalar variable x with a scalar, say, α can be similarly defined as

$$(\alpha\mathbf{A})(x) = \alpha\mathbf{A}(x). \quad (2-50)$$

Other algebraic operations with vector functions can be similarly defined by first referring to an arbitrarily chosen value of the scalar variable x , and then letting x vary over the relevant interval.

2.11.1 The derivative of a vector function

The derivative of a scalar function $A(x)$ of a scalar variable x at the point x is defined by the usual limiting expression

$$\frac{dA}{dx} = \lim_{h \rightarrow 0} \frac{A(x+h) - A(x)}{h}. \quad (2-51)$$

Making use of this definition, one can define the derivative of a *vector* function $\mathbf{A}(x)$ in terms of the derivatives of the three scalar functions $A_i(x)$ ($i = 1, 2, 3$) corresponding to the three scalar components of $\mathbf{A}(x)$ as

$$\frac{d\mathbf{A}}{dx} = \frac{dA_1}{dx}\hat{i} + \frac{dA_2}{dx}\hat{j} + \frac{dA_3}{dx}\hat{k}. \quad (2-52)$$

For instance, the derivative of the vector function expressed by eq. (2-47) for any given value of x is

$$\frac{d\mathbf{A}}{dx}(x) = \cos x \hat{i} - \sin x \hat{j} - 2x e^{-x^2} \hat{k}. \quad (2-53)$$

Problem 2-15

Let $\mathbf{A}(t)$, and $\mathbf{B}(t)$ be two given vector functions of the scalar variable t . Consider the scalar function $\mathbf{A}(t) \cdot \mathbf{B}(t)$ of t . Show that

$$\frac{d}{dt}(\mathbf{A}(t) \cdot \mathbf{B}(t)) = \frac{d\mathbf{A}(t)}{dt} \cdot \mathbf{B}(t) + \frac{d\mathbf{B}(t)}{dt} \cdot \mathbf{A}(t). \quad (2-54)$$

Answer to Problem 2-15

HINT: Express each of the two vector functions $\mathbf{A}(t)$, $\mathbf{B}(t)$ in terms of three scalar component functions with respect to any chosen Cartesian co-ordinate system. However, the result can be established even without referring to a co-ordinate system. Try it out.

Problem 2-16

For a vector function $\mathbf{A}(t)$ of time represented by the scalar variable t , show that

$$\frac{d}{dt}A(t)^2 = 2\mathbf{A}(t) \cdot \frac{d\mathbf{A}(t)}{dt}. \quad (2-55)$$

Answer to Problem 2-16

HINT: The squared magnitude of the vector at time t is given by $(A(t))^2 = \mathbf{A}(t) \cdot \mathbf{A}(t)$; now apply the result of eq. (2-54).

2.12 Scalar function of a vector variable

2.12.1 The position vector

Consider a region R in space (fig. 2-19) and an origin O that may or may not lie within R . Given any point P in R , the vector represented by the directed line segment extending from O to P is referred to as the *position vector* of P with respect to O . One can thus identify each point in R with a help of a position vector, where a position vector corresponds to a uniquely defined point.

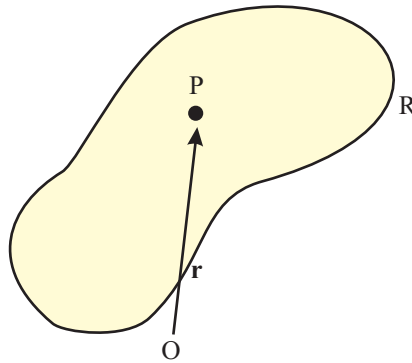


Figure 2-19: Illustrating the idea of the position vector of a point in a given region R ; the position vector depends on the choice of the origin (O); for a given origin O , the position vector of the point P is the vector represented by the directed line segment extending from O to P ; the position vector may be used to identify points within the region R .

Choosing a Cartesian co-ordinate system with O as the origin, one can represent the position vector \mathbf{r} of the point P in terms of the Cartesian co-ordinates x, y, z of P :

$$\mathbf{r} = x\hat{i} + y\hat{j} + z\hat{k}. \quad (2-56)$$

The region R may even be imagined to extend through entire space, in which case, each

point of space is defined in terms of a position vector \mathbf{r} (or, equivalently, of the Cartesian co-ordinates x, y, z) with reference to the origin O (and to the co-ordinate system chosen).

At times, a phrase like ‘the point with the position vector \mathbf{r} ’ is shortened to one sounding like ‘the point \mathbf{r} ’, where the origin may or may not be explicitly referred to.

Problem 2-17

Consider three points P, Q, R with position vectors, respectively, $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3$, with reference to some chosen origin O . Obtain the condition for the points to be collinear.

Answer to Problem 2-17

HINT: Recall that, given any set of three points, a plane can always be drawn containing the points (in the special case of the points being collinear, an *infinite* number of such planes exist; reason this out). The condition of collinearity is that the vector $\mathbf{r}_3 - \mathbf{r}_1$ is to be parallel to $\mathbf{r}_3 - \mathbf{r}_2$, i.e., the cross product of these two vectors is to be zero (reason this out), in which case the triangle with the three points as its vertices will be of area zero (the triangle collapses to a straight line). The required condition reads

$$\mathbf{r}_1 \times \mathbf{r}_2 + \mathbf{r}_2 \times \mathbf{r}_3 + \mathbf{r}_3 \times \mathbf{r}_1 = 0. \quad (2-57)$$

Problem 2-18

The parametric equation of a straight line passing through a given reference point \mathbf{r}_0 and parallel to the a given unit vector \hat{n} is of the form

$$\mathbf{r} = \mathbf{r}_0 + \lambda \hat{n}, \quad (2-58)$$

where λ denotes the signed distance of the variable point \mathbf{r} from the reference point \mathbf{r}_0 (reason this statement out). Consider two lines L_1, L_2 with reference points $\mathbf{r}_1, \mathbf{r}_2$ (these may be taken to be distinct without loss of generality) and unit vectors \hat{n}_1, \hat{n}_2 . Obtain the condition of the two being (a) parallel, and (b) skew to each other (i.e., they do not lie in a single plane; a third possibility is that the two lines lie in a single plane, and intersect each other).

Answer to Problem 2-18

HINT: (a) $\hat{n}_1 \times \hat{n}_2 = 0$. (b) In this case $\hat{n}_1 \times \hat{n}_2 \neq 0$. If $\hat{n}_1 \times \hat{n}_2$ is perpendicular to $\mathbf{r}_2 - \mathbf{r}_1$ (which is non-zero by assumption) then the two lines are co-planar and have a common point. Thus, the required condition is

$$(\hat{n}_1 \times \hat{n}_2) \cdot (\mathbf{r}_2 - \mathbf{r}_1) \neq 0. \quad (2-59)$$

2.12.2 Scalar function of the position vector: scalar field

Imagine a region R in space in which a scalar ϕ is associated uniquely with every point with position vector \mathbf{r} in the region. This correspondence between \mathbf{r} and the scalar ϕ can be expressed symbolically by writing ϕ as $\phi(\mathbf{r})$, and is said to define a scalar *field* ϕ in the given region. Generally speaking, the term ‘field’ is used to indicate some function of the position variable \mathbf{r} in some given region of space (there are, however, other connotations of the term ‘field’; for instance a field may mean a set having a number of specified characteristics, the elements of which are termed *scalars* with reference to a linear vector space).

As an example, consider the region of space within a chamber in which the temperature T may vary from one point to another. At any given instant of time, the temperature has some specific value at any given point with position vector \mathbf{r} within the chamber. One then has a scalar field T , for which the symbol $T(\mathbf{r})$ may be used to denote the temperature at the point \mathbf{r} .

Representing the position vector \mathbf{r} of a point in terms of the Cartesian co-ordinates (x, y, z) of any chosen co-ordinate system, one can look upon a scalar field as being a scalar function of three scalar variables, and express $\phi(\mathbf{r})$ as $\phi(x, y, z)$.

For instance, the expression

$$\phi(\mathbf{r}) = \sin x e^{-y^2} \cos z, \quad (2-60)$$

defines a scalar field for each of the co-ordinates x, y, z varying from $-\infty$ to ∞ .

More generally, given a vector variable \mathbf{v} , corresponding to three scalar variables v_1, v_2, v_3 representing the components of \mathbf{v} , one can define a scalar function ϕ of \mathbf{v} , where an alternative description of ϕ is to express it as a function of the three scalar variables, in the form $\phi(v_1, v_2, v_3)$. For instance, the speed (c) of a particle is a scalar function of its velocity vector \mathbf{v} or equivalently, of its velocity components v_1, v_2, v_3 , each of which can vary in the range $-\infty$ to ∞ :

$$c = (\mathbf{v} \cdot \mathbf{v})^{\frac{1}{2}} = (v_1^2 + v_2^2 + v_3^2)^{\frac{1}{2}}. \quad (2-61)$$

In a manner of speaking, this may be looked upon as defining a scalar field in some region of a 'space' made up of the co-ordinates v_1, v_2, v_3 .

2.13 Vector function of a vector variable: vector field

Imagine a region R in space such that, given any point \mathbf{r} in this region, there is associated a uniquely specified vector that we denote by the symbol $\mathbf{v}(\mathbf{r})$. This association of $\mathbf{v}(\mathbf{r})$ with \mathbf{r} defines a *vector function* of the position vector \mathbf{r} or, equivalently, a *vector field* in the region R . Referring to any chosen Cartesian co-ordinate system, the vector field can be described in terms of three scalar fields, each corresponding to some component of \mathbf{v} :

$$\mathbf{v}(\mathbf{r}) = v_1(x, y, z)\hat{i} + v_2(x, y, z)\hat{j} + v_3(x, y, z)\hat{k}. \quad (2-62)$$

For instance, the expression

$$\mathbf{v}(\mathbf{r}) = e^{-\mathbf{r} \cdot \mathbf{r}}\hat{i} + e^{-2\mathbf{r} \cdot \mathbf{r}}\hat{j} + e^{-3\mathbf{r} \cdot \mathbf{r}}\hat{k}, \quad (2-63)$$

defines a vector field where R may be taken to be extended throughout space.

As an example, imagine the region R covering the interior of a room through which a

draft of air is blowing. At any given instant of time, the mean velocity of the air molecules around any given point \mathbf{r} can be specified in terms of $\mathbf{r}(x, y, z)$ in the form of a vector function $\mathbf{v}(\mathbf{r})$, thereby giving the velocity field in the room.

One can represent a vector field \mathbf{v} *geometrically* by imagining a directed line segment representing the vector $\mathbf{v}(\mathbf{r})$ attached to every point \mathbf{r} of the region R . Fig. 2-20 depicts such a vector field with the arrows representing the vectors associated with the initial points of the directed line segments.

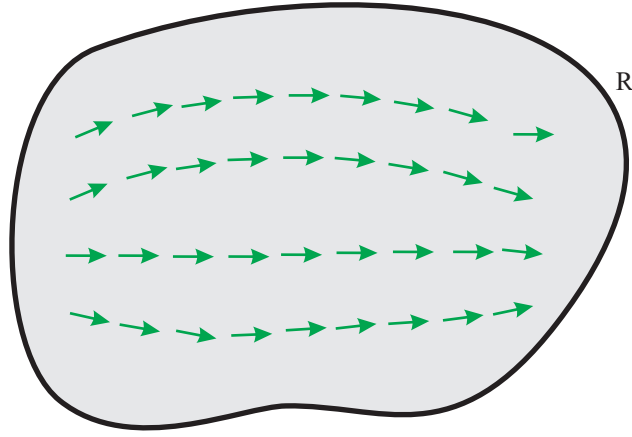


Figure 2-20: Illustrating a vector field in a region R (schematic); the vector field is represented by a directed line segment attached to every point \mathbf{r} in R and corresponds to a vector function of the position vector \mathbf{r} .

2.14 Derivatives and integrals of scalar and vector fields

2.14.1 Derivatives of scalar and vector fields

The idea of the derivative can be extended to scalar and vector fields in a straightforward manner. For instance, given a scalar field $\phi(x, y, z)$, one can define the *partial derivatives* $\frac{\partial \phi}{\partial x}$, $\frac{\partial \phi}{\partial y}$, $\frac{\partial \phi}{\partial z}$ at any given point (x, y, z) . Here a partial derivative such as, say, $\frac{\partial \phi}{\partial x}$ is one obtained by taking the derivative of ϕ with respect to x , with the remaining variables y, z *held constant*. For instance, for the scalar field ϕ defined by the expression (2-60), one has $\frac{\partial \phi}{\partial x} = \cos x \, e^{-y^2} \cos z$.

Making use of the three partial derivatives of a scalar field ϕ with respect to x , y , and z , one can construct a vector field \mathbf{v} with the help of the expression

$$\mathbf{v}(\mathbf{r}) = \frac{\partial \phi}{\partial x} \hat{i} + \frac{\partial \phi}{\partial y} \hat{j} + \frac{\partial \phi}{\partial z} \hat{k}. \quad (2-64)$$

This vector field holds a special significance with reference to the scalar field ϕ and is termed the *gradient* of ϕ .

Considering, on the other hand, a vector field \mathbf{v} with Cartesian components v_1 , v_2 , v_3 forming three scalar fields, one may similarly form scalar and vector fields with the help of the partial derivatives of these components with respect to x, y, z . One such scalar field, having a special significance with reference to the vector field \mathbf{v} , is its *divergence* defined as

$$\text{div } \mathbf{v}(\mathbf{r}) = \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} + \frac{\partial v_3}{\partial z}. \quad (2-65)$$

Another vector field defined with reference to the partial derivatives of the components of $\mathbf{v}(\mathbf{r})$ is its *curl*, defined as

$$\text{curl } \mathbf{v}(\mathbf{r}) = \left(\frac{\partial v_3}{\partial y} - \frac{\partial v_2}{\partial z} \right) \hat{i} + \left(\frac{\partial v_1}{\partial z} - \frac{\partial v_3}{\partial x} \right) \hat{j} + \left(\frac{\partial v_2}{\partial x} - \frac{\partial v_1}{\partial y} \right) \hat{k}. \quad (2-66)$$

Problem 2-19

Consider the special vector field $\mathbf{v}(\mathbf{r}) = \mathbf{r}$, i.e., the one that associates with each point \mathbf{r} , the vector \mathbf{r} itself. For this vector field, which we denote by the symbol \mathbf{r} for the sake of brevity, work out (a) *div* \mathbf{r} , *grad* $|\mathbf{r}|^2$, and *curl* \mathbf{r} .

Answer to Problem 2-19

HINT: Note that the partial derivatives of the components of \mathbf{r} with respect to the co-ordinates are of the form $\frac{\partial x}{\partial x} = 1$, $\frac{\partial x}{\partial y} = 0$, etc. Making use of these and the defining relations (2-64), (2-65), (2-66), along with the relation $|\mathbf{r}|^2 = \mathbf{x}^2 + \mathbf{y}^2 + \mathbf{z}^2$, show that (a) *div* $\mathbf{r} = 3$, (b) *grad* $|\mathbf{r}|^2 = 2\mathbf{r}$, (c) *curl* $\mathbf{r} = 0$. These are useful relations in vector calculus.

2.14.2 Volume-, surface-, and line integrals

The basic idea underlying the definition of volume-, surface-, and line integrals involving scalar and vector fields relates to the fundamental principle of integral calculus that expresses an integral as the limit of a sum. For instance, consider a volume V in a region R of space. Imagine the volume V to be partitioned into a large number of small volume elements, a typical volume element around the point \mathbf{r} being, say δV (fig. 2-21). If, now ϕ denotes a scalar field defined throughout the region R , one can consider a sum of terms like $\phi(\mathbf{r})\delta V$, where the summation is taken over all the volume elements making up the volume V . Assuming that all the volume elements δV are infinitesimally small (i.e., tend to the limit zero), the above sum gives the *volume integral* of ϕ over the volume V :

$$\int_V \phi dv = \lim_{\{\delta V\} \rightarrow 0} \sum \phi(\mathbf{r})\delta V. \quad (2-67)$$

This volume integral corresponds to a scalar determined by the scalar field ϕ and the volume of integration V .

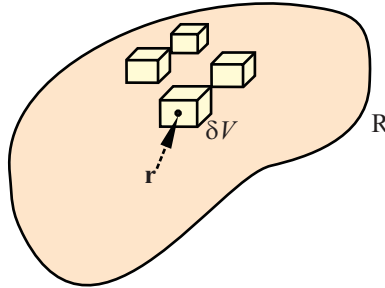


Figure 2-21: Illustrating the idea of the volume integral; the volume V in the region R is imagined to be divided into a large number of small volume elements, a number of which are shown schematically, one such element being δV around the point \mathbf{r} ; the volume integral of a scalar field $\phi(\mathbf{r})$ is then defined as a sum of terms like $\phi(\mathbf{r})\delta V$, over all these volume elements, in the limit when $\delta V \rightarrow 0$ for each of the elements.

Similarly, consider a surface S within a given region R and imagine the surface to be partitioned into a large number of small surface elements, where each surface element can effectively be considered to be a planar surface. Let the vector $\delta S \hat{n}$ denote the vector

area of a surface element around any given point \mathbf{r} , where δS denotes the magnitude of the surface area and \hat{n} denotes the unit normal to the surface at the point \mathbf{r} , any one of the two possible orientations of the normal being chosen for the purpose of defining the surface integral (fig. 2-22(A)), the *same* orientation relative to the surface being chosen for all the surface elements under consideration.

Consider now a vector field \mathbf{A} defined throughout the region R , including the points on the surface S . For this vector field \mathbf{A} and this surface S , consider a sum of terms of the form $\mathbf{A}(\mathbf{r}) \cdot \hat{n} \delta S$, where the summation is to be carried out over all the surface elements making up the surface S . Assuming now that each of these surface elements is of an infinitesimally small area ($\delta S \rightarrow 0$), the limiting value of the above sum is termed the *surface integral* of the vector field (or, in brief, of the vector) \mathbf{A} over the surface S , relative to the chosen orientation of the normals at various points on S :

$$\int_S \mathbf{A} \cdot \hat{n} dS = \lim_{\{\delta S\} \rightarrow 0} \sum \mathbf{A} \cdot \hat{n} \delta S. \quad (2-68)$$

Consider, as a special case, a *closed* surface S . In this case, an orientation of the normals at the various points can be chosen by referring to the *outward* or the *inward* direction relative to the volume enclosed by the surface. A common choice for \hat{n} in such a case is the outward drawn normal to S at any given point (fig. 2-22(B)). One thereby arrives at the *closed* surface integral expressed as $\oint \mathbf{A} \cdot \hat{n} dS$.

Finally, consider a line or a *path* in a region R in which a vector field \mathbf{A} is defined, where the path connects any two given points P and Q (fig. 2-23(A)). The path may be imagined to be made up of a large number of small segments, where each segment may be assumed to be effectively a linear one. Considering any particular sense of traversal of the path, say, from P to Q , let a typical segment around any given point, say \mathbf{r} , on the path be of length δl . One can represent the vector length of this segment by a vector $\delta \mathbf{r}$, whose magnitude is δl and which is directed from the initial to the final point from the segment (relative to the sense of traversal of the path).

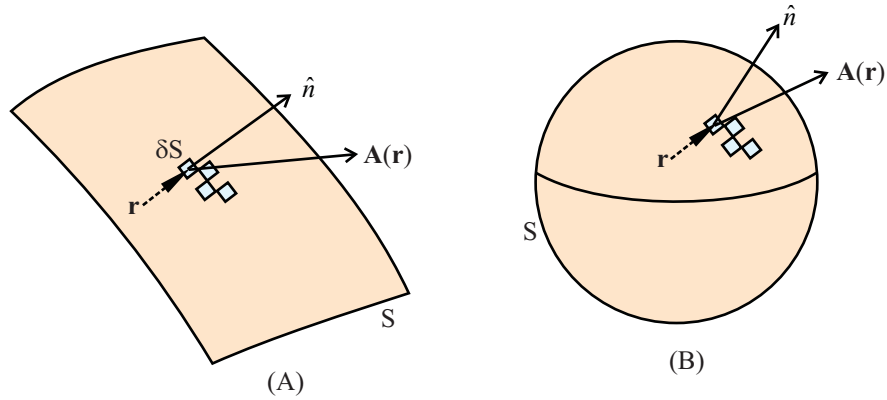


Figure 2-22: Illustrating the idea of a surface integral of a vector field; the surface S lying in some region R (not shown) is imagined to be partitioned into a large number of infinitesimally small surface elements, a number of such elements being shown schematically for (A) an open surface, and (B) a closed surface; the unit normal to a typical element around any point \mathbf{r} on the surface is denoted by $\hat{\mathbf{n}}$, where any one of the two possible orientations for $\hat{\mathbf{n}}$ may be chosen; for a closed surface, however, the *outward* normal is commonly chosen; the surface integral is a sum of terms of the form $\mathbf{A} \cdot \hat{\mathbf{n}} \delta S$, taken over all the elements making up the surface, in the limit $\delta S \rightarrow 0$ for all the surface elements.

Consider now a sum of terms like $\mathbf{A}(\mathbf{r}) \cdot \delta \mathbf{r}$, where the summation is to be taken over all the segments making up the path from P to Q . Assuming that the lengths δl of all these segments are vanishingly small, the sum gives us the *line integral* (also termed a *path integral*) of the vector field (or, in brief, of the vector) \mathbf{A} along the path:

$$\int_P^Q \mathbf{A} \cdot d\mathbf{r} = \lim_{\{\delta l\} \rightarrow 0} \sum \mathbf{A}(\mathbf{r}) \cdot \delta \mathbf{r}. \quad (2-69)$$

In particular, considering a *closed* path for which the points P and Q are the same, and choosing a sense of traversal of the path, the limiting value of the above sum gives the closed line integral of \mathbf{A} , denoted by the symbol $\oint \mathbf{A} \cdot d\mathbf{r}$, with reference to the chosen sense of traversal (fig. 2-23(B)).

The line integral is sometimes denoted in the alternative forms

$$\int \mathbf{A} \cdot d\mathbf{r} = \int \mathbf{A}(\mathbf{r}) \cdot \vec{dl} = \int \mathbf{A} \cdot \hat{\mathbf{t}} dl, \quad (2-70)$$

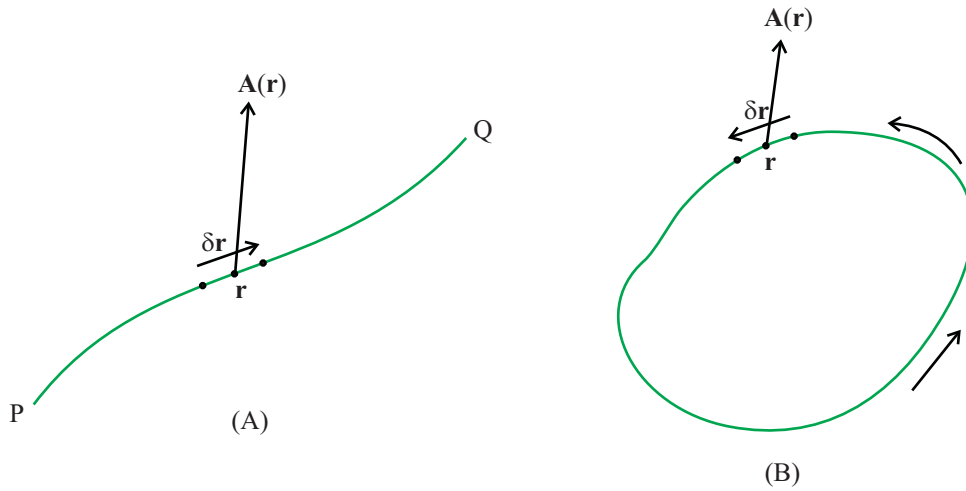


Figure 2-23: Illustrating the idea of the line integral of a vector field; (A) an open path connecting the points P and Q, (B) a closed path; the sense of traversal in (A) is from P to Q (say), while that in (B) is indicated by the bent arrows; the path is imagined to be divided into a large number of vanishingly small segments where each segment is effectively a part of a straight line; the line integral is then defined as a sum over all these segments; $\delta \mathbf{r}$ denotes the vector length of a typical segment around any point \mathbf{r} on the path.

where $\vec{dl}(= \lim \delta \mathbf{r})$ denotes the directed line segment extending from the initial to the final point of a typical infinitesimal line segment around the point \mathbf{r} , and \hat{t} denotes the unit vector along the tangent to the path at the point \mathbf{r} , directed in the sense of traversal of the path under consideration. Usually the notation and the meanings of symbols are to be understood from the context.

Problem 2-20

Consider the closed path formed by the line segments OP, PQ, and QO in the x-y plane, where O is the origin, P is the point with co-ordinates (1, 0) and Q is the point (0, 1) relative to a Cartesian co-ordinate system. Consider further the vector field $\mathbf{A}(x, y, z) = y\hat{i} + x\hat{j} + xy\hat{k}$. Work out the line integral of \mathbf{A} along the closed path traversed in the sense OPQO.

Answer to Problem 2-20

The required line integral breaks up into three terms, one each for the segments OP, PQ, and QO. Consider, for instance, the segment PQ, for which we have $z = 0, y = 1 - x$. Thus the typical vector line element on this segment will be of the form $d\mathbf{r} = dx\hat{i} + dy\hat{j} = (\hat{i} - \hat{j})dx$ while, on this segment,

$A(x, y, z) = (1 - x)\hat{i} + x\hat{j} + x(1 - x)\hat{k}$. Evaluating the scalar product $\mathbf{A} \cdot d\mathbf{r}$ the line integral on this segment is found to be $\int_1^0 (1 - 2x)dx = 0$. The contributions from the other two segments can be similarly worked out, giving $\oint \mathbf{A} \cdot d\mathbf{r} = 0$ (the integrals along the other two segments also vanish).

Problem 2-21

Consider the vector field $\mathbf{v}(\mathbf{r}) = \mathbf{r}$ and the sphere centered at the origin, having a radius a . Obtain (a) the volume integral of the divergence of the vector field over the volume V of the sphere, and (b) the surface integral of the vector field over the surface S of the sphere.

Answer to Problem 2-21

HINT: As seen in problem 2-19, $\text{div}\mathbf{r} = 3$, and thus, (a) $\int_V \text{div}\mathbf{r} dV = 3 \int_V dv$, which is 3 times the volume of the sphere (arrived at by summing all the small volume elements δV), i.e., $4\pi a^3$. In (b), on the other hand, we have to sum up terms like $\mathbf{r} \cdot \hat{n} \delta S$ over the surface of the sphere, where \hat{n} denotes the outward drawn normal at any chosen point on the surface. In this expression, one can put $\mathbf{r} = a\hat{n}$ (reason out why), which gives, for the surface integral, a sum of terms of the form $a\delta S$, i.e., a times the sum of all the surface elements or, finally, $4\pi a^3$.

Note how the two answers in parts (a) and (b) of problem 2-21 turn out to be the same. This is a particular instance of a more general result, referred to as *Gauss' theorem* in vector calculus, which states that the volume integral of the divergence of a vector field $\mathbf{A}(\mathbf{r})$ over a volume V equals the surface integral (evaluated in the sense of the outward drawn normal) of the vector field over the closed surface S enclosing the volume:

$$\int_V \text{div} \mathbf{A}(\mathbf{r}) dV = \oint_S \mathbf{A}(\mathbf{r}) \cdot \hat{n} dS. \quad (2-71)$$

Gauss' theorem in vector calculus is a remarkably useful one. It forms the basis of Gauss' principle in gravitation (refer to section 5.3) and of Gauss' principle in electrostatics (see section 11.8), the two principles being closely similar in content.

Chapter 3

Mechanics

3.1 Introduction: frames of reference

While performing a measurement on some observed system, an observer usually employs a set of measuring apparatus *stationary* with respect to herself. A second observer, performing observations and measurements of her own, also makes use of apparatus stationary with respect to herself. One expresses this by saying that the two observers are making measurements in their own respective *frames of reference*. If the second observer happens to be stationary with respect to the first, then one can say that their sets of measuring apparatus effectively belong to the *same* frame of reference. If, however, the two observers have a relative motion, then their frames of reference will be distinct from each other.

In other words, while doing physics one has to keep in mind that every measurement is in reality performed in some particular frame of reference or other, and it does not pay to confuse one frame of reference with another. All objects rigidly fixed with respect to one another constitute a frame of reference. However, one need not always refer to a set of objects in defining a frame of reference, since a set of imagined lines and points, rigidly fixed with respect to one another, may also serve the purpose (though, in practice, such lines and points are also defined with reference to objects and entities existing around us).

Suppose that the straight lines OX , OY , and OZ (fig. 3-1) are rigidly attached with one another. Then these three lines specify a frame of reference. However, if OX' , OY' , OZ' happen to be three other straight lines rigidly fixed with one another and also stationary with respect to OX , OY , OZ , then these three correspond to the *same* frame of reference. One then says that OX , OY , OZ , and OX' , OY' , OZ' constitute two different *co-ordinate systems* in the same frame of reference. In other words, a co-ordinate system specifies a frame of reference, but the same frame of reference may accommodate an infinite number of other co-ordinate systems as well. Two such co-ordinate systems have been shown in fig. 3-1. In the following, I shall often use the term frame of reference while referring to some particular co-ordinate system in it.

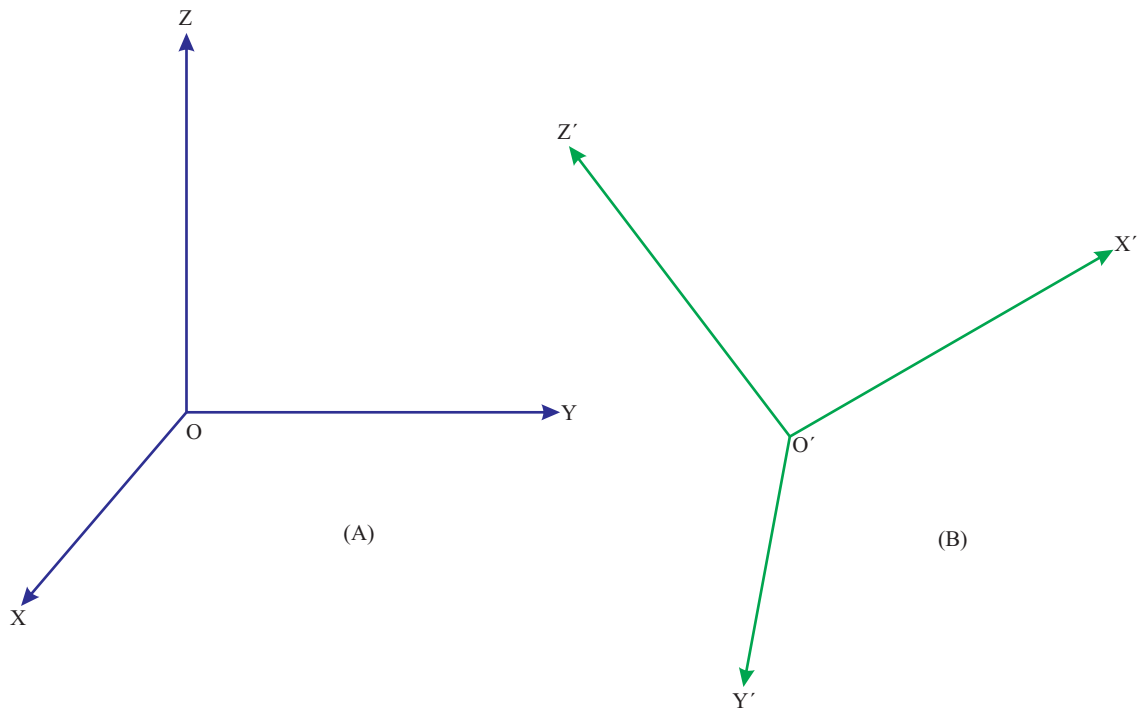


Figure 3-1: Two co-ordinate systems stationary with respect to each other; (A) a right handed Cartesian co-ordinate system, and (B) a left handed Cartesian co-ordinate system; both the co-ordinate systems are in the same frame of reference.

Classification of co-ordinate systems

Both the co-ordinate systems shown in fig. 3-1 are *Cartesian*, or *rectilinear orthogonal*

systems, because in each of these, the axes are straight lines perpendicular to one another. However, not all co-ordinate systems are of the rectilinear orthogonal type. For instance, a co-ordinate system may be made up of three non-coplanar straight lines obliquely inclined to one another, constituting a rectilinear *oblique* co-ordinate system. Co-ordinate systems may also be made up of *curved* lines. *Curvilinear orthogonal* co-ordinate systems belong to this class, examples being the *spherical polar* and *cylindrical polar systems*.

In this book we shall mostly refer to rectilinear orthogonal co-ordinate systems. Among these, one further distinguishes between *right handed* and left handed systems. Of the two systems shown in fig. 3-1, (A) depicts a right handed system while (B) depicts a left handed one.

In chapter 2, I introduced left handed and right handed orthonormal triads of vectors and mentioned that a Cartesian co-ordinate system has associated with it an orthonormal triad. Notice that the unit vectors (say, \hat{i} , \hat{j} , \hat{k}) along the three axes (respectively OX, OY, OZ) of the system in fig. 3-1(A) form a right handed triad, while the corresponding unit vectors in (B) form a left handed one. In other words, the definition of left handed and right handed Cartesian co-ordinate systems corresponds to that of left handed and right handed orthonormal triads.

Considering two right handed (or two left handed) Cartesian co-ordinate systems in the same frame of reference, one can be made to coincide with the other by processes of parallel translation and rotation. However, a right handed system *cannot* be made to coincide with a left handed one by parallel translation and rotation. For this, one needs to reverse the direction of one axis (or all three axes) of one of the systems. This process is termed an axis *inversion*. While the inversion of an odd number of axes changes the handedness of a co-ordinate system, an inversion of an even number of axes leaves the handedness unaltered.

3.2 Motion of a single particle

3.2.1 Introduction

Dynamics is a branch of mathematics and physics in which one aims to describe the motion of any system made up of one or more bodies, and to explain such motion in *causal* terms. In this, the *descriptive* aspect is sometimes identified as a separate branch called *kinematics*, while the causal explanation is identified as dynamics proper.

Among the various different types of systems studied in kinematics and dynamics, the simplest is a system made up of just one single particle.

The concept of a particle is, however, an *idealized* one. One assumes that a particle possesses a *mass* but does not have a spatial extension.

As in the case of observations and measurements in general, one has to refer to some particular frame of reference or other in order to describe and explain the motion of a single particle. The relations between the relevant physical quantities obtained in the course of such description and explanation are then specific to that frame of reference.

For a single particle, one requires at the outset *three* independent physical quantities so as to specify its position. These one may choose as the three co-ordinates of the instantaneous position of the particle relative to any Cartesian co-ordinate system (commonly chosen to be a right handed one) in the given frame of reference. These are measured with the help of appropriate measuring rods. Their values are denoted by the symbols x , y , and z . One other physical quantity of relevance is the *time*, denoted by the symbol t and measured with the help of an appropriate clock.

In this context, it is important to distinguish between *relativistic* and *non-relativistic* points of view in mechanics. A fundamental idea in the relativistic theory is that the clock used by an observer in measuring time is an integral part of her frame of reference. Among all the clocks stationary with respect to the observer, she may choose any one for the sake of time measurement. If, along with this, she chooses a co-ordinate system of her own, then that will determine her frame of reference.

In this frame, the fact that the particle is at some definite position at any given time instant, is completely described by the four quantities, t , and x, y, z , where these four taken together are said to specify an *elementary event*. The same elementary event will be described by some other observer, moving with respect to the former, by another set of quantities, say, t' and x', y', z' , obtained by measurements in a second frame of reference. The clock used in this case has to be included in the specification of this second frame.

Even if the readings of the clocks used in the two frames be made to agree for some particular chosen event, these will be found to differ, in general, for *other* elementary events observed from the two frames. In other words, in the relativistic theory, the measured value of time differs in different frames : two clocks in motion relative to each other correspond to two different frames of reference (see fig. 3-2; you will find a brief exposition of the basic ideas involved in the relativistic theory in chapter 17). This is what is referred to as *relativity of time*.

By contrast, the measurement of time is *independent* of the observer in the non-relativistic theory. In this theory, it is assumed that once the readings of two clocks in relative motion are made to agree, they will continue to agree at all times. Stated another way, the choice of a clock is not essential in specifying a frame of reference. Even if the clocks for two observers in relative motion are not made to agree at any given time, it does not really matter, because the *difference* in their readings will continue to remain the same. In other words, the *time-interval* between any two events will be the same for the two clocks.

In the present state of our knowledge based on observations, one can state that, of these two points of view, the relativistic approach is the more correct and acceptable one. However, the non-relativistic theory is also a useful and competent one, though within certain limits of observation and experience. If the relative velocity between two observers is small (compared to the *velocity of light in vacuum*) then the difference between the time-intervals measured by their respective clocks can, to a certain degree of approximation, be ignored.

With this limitation in mind, I shall, in this chapter, present a brief outline of the

basics of non-relativistic kinematics and dynamics, where spatial co-ordinates (as also velocities, see sec. 3.9.1) are, in general, frame dependent, but time-intervals are frame independent.

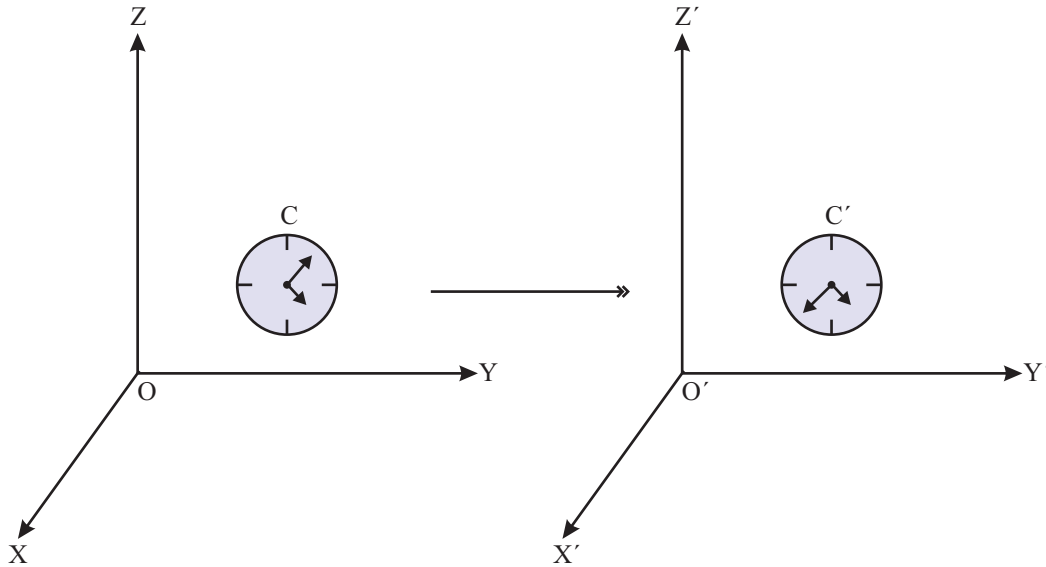


Figure 3-2: Two frames of reference in relative motion (indicated by double-headed arrow); each frame includes a co-ordinate system *along with* a clock of its own (clock C for co-ordinate system OXYZ and C' for O'X'Y'Z'); time-intervals indicated by the two clocks need not be the same.

3.2.2 Kinematic quantities

I have already mentioned that in the description of motion of a single particle in any given frame of reference, the physical quantities to start with are the position co-ordinates x , y , z relative to any chosen Cartesian co-ordinate system in that frame, and the time t .

In general, the position of the particle changes with time in accordance with certain *causal* rules, determined by the way the particle interacts with *other* particles or bodies. Depending on the nature of this interaction, the position co-ordinates x , y , z change with time t in different manners in different specific situations. It is the aim of *dynamics* to determine the nature of this variation with time, but I shall come to that later. What is of interest now is that, in any given situation, x , y , and z can be considered to be

some *functions* of time t , as a result of which these can be written as, say, $x(t)$, $y(t)$, and $z(t)$.

3.2.2.1 The position vector and its time dependence

Consider the co-ordinate system in terms of which the position co-ordinates x , y , and z have been defined. The vector extending from the origin O of this system to the instantaneous position, say, P of the particle under consideration, will be referred to as the *position vector* of the particle (see fig. 3-3, where P denotes the position at any given instant, say, t). Denoting this as \mathbf{r} , the variation of \mathbf{r} with time t due to the motion of the particle, can be represented by expressing \mathbf{r} as a *vector function* of t . Writing this function as $\mathbf{r}(t)$, one has

$$\mathbf{r}(t) = x(t)\hat{i} + y(t)\hat{j} + z(t)\hat{k}, \quad (3-1)$$

where \hat{i} , \hat{j} , \hat{k} are the unit vectors associated with the chosen co-ordinate system.

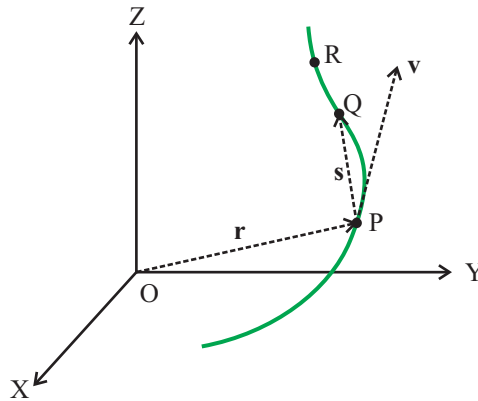


Figure 3-3: Possible trajectory of a single particle; P , Q , R denote positions of the particle at three successive instants of time; the position vector (\mathbf{r}) at the position P is represented by the directed line segment extending from the origin O of the co-ordinate system, up to P ; \mathbf{s} denotes the displacement from P (time t) to Q (time $t + T$); \mathbf{v} , the velocity of the particle at P , is along the tangent to the trajectory, drawn at P ; the magnitude of this vector gives the speed at P .

The concept of the position vector was introduced earlier in sec. 2.12.1. The description of the motion of a particle consists of specifying how its position vector \mathbf{r} changes

with time t .

3.2.2.2 Displacement

Fig. 3-3 depicts the possible trajectory of a single particle where P and Q denote the positions of the particle at times t and $t + T$ (say), the time interval between the two being T . The position vectors at these two instants may be denoted by $\vec{OP} = \mathbf{r}(t)$ and $\vec{OQ} = \mathbf{r}(t + T)$ respectively, where O is the origin of the co-ordinate system (OXYZ) with respect to which the components of various vectors will be indicated. The directed line segment extending from P to Q then represents what is termed the *displacement* (say, \mathbf{s}) from P to Q, occurring in the time interval T . One thus has

$$\mathbf{s} = \mathbf{r}(t + T) - \mathbf{r}(t). \quad (3-2)$$

If the co-ordinates of the particle at times t and $t + T$ be (x_0, y_0, z_0) , and (x, y, z) respectively, then one can write, in terms of these co-ordinates,

$$\mathbf{s} = (x - x_0)\hat{i} + (y - y_0)\hat{j} + (z - z_0)\hat{k}. \quad (3-3)$$

In special situations the particle under consideration may be *at rest* for some given interval of time. In such a situation, $\mathbf{r}(t)$ will remain unchanged during that interval, and the corresponding displacement will be zero (the null vector). However, when looked at from some *other* frame of reference, the particle may well be in motion and suffer a displacement during the interval under consideration. In other words, rest and motion are *relative* concepts, depending on the frame of reference from which the particle is observed. Indeed, the description of motion in one frame differs, in general, from that in another. Two such descriptions, however, cannot be unrelated to each other since both refer to the motion of the *same* particle. Their relation is expressed in terms of a set of *transformation formulae* between the two frames of reference. Later in this chapter we will come across instances of such transformation formulae. Transformation formulae are also needed in relating the descriptions of the motion in two co-ordinate systems in the same frame of reference, as seen in the following problem.

Problem 3-1

Consider the co-ordinate systems S, S' of fig. 2-17, stationary with respect to each other (i.e., both in the same frame of reference), where S' is obtained from S by a rotation about the axis OZ through an angle $\phi = \frac{\pi}{3}$. The time-dependent position vector of a moving particle is given by $\mathbf{r}(t) = t^2 \hat{n}$, where the unit vector \hat{n} has direction cosines $\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}$ with reference to the three axes in S. Obtain the expression for $\mathbf{r}(t)$ in terms of the unit vectors $\hat{i}', \hat{j}', \hat{k}'$ along the three axes of S'.

Answer to Problem 3-1

HINT: Referring to the transformation matrix T of problem 2-11 with $\phi = \frac{\pi}{3}$, giving the transformation of the components of a vector in S to those in S', and noting that the Cartesian components of a unit vector give its direction cosines (refer to sec. 2.5.4), one obtains, $\mathbf{r}(t) = t^2 \left\{ \frac{1}{2} \left(1 + \frac{1}{\sqrt{3}} \right) \hat{i}' + \frac{1}{2} \left(-1 + \frac{1}{\sqrt{3}} \right) \hat{j}' + \frac{1}{\sqrt{3}} \hat{k}' \right\}$.

3.2.2.3 Velocity

The displacement being s in a time-interval T , the *average rate* of displacement in this interval is $\frac{s}{T}$. In general, if one divides the interval T into smaller sub-intervals, then the rates of displacement during these intervals will be found to differ from one another as also from the above mean rate of displacement. However, if the interval T is a sufficiently *small* one, then the rates of displacement during the sub-intervals will be found to be close to one another. Notice that as the interval T is imagined to be made progressively smaller, the point Q in the trajectory will gradually approach P and in the end, the vector $s = \vec{PQ}$, of infinitesimally small length, will point in the direction of the *tangent* at P (fig. 3-3). The rate of displacement in the limit $T \rightarrow 0$ is termed the *velocity* of the particle at time t (i.e., at the position P). Denoting this as \mathbf{v} , one gets the mathematical definition

$$\mathbf{v} = \lim_{T \rightarrow 0} \frac{s}{T}. \quad (3-4)$$

Evidently, the velocity \mathbf{v} will be a vector directed along the tangent to the trajectory at

P. The magnitude of this vector is given by

$$|\mathbf{v}| = \lim_{T \rightarrow 0} \frac{|\mathbf{s}|}{T}. \quad (3-5)$$

As a special case, if the displacement takes place *at a uniform rate* during some interval T , then the limit $T \rightarrow 0$ will no longer be necessary and one can write, simply, $\mathbf{v} = \frac{\mathbf{s}}{T}$. In this special case of *uniform motion*, the trajectory of the particle will be a straight line and the velocity will be a vector directed along this straight line. Though the velocity is independent of time for a particle in uniform motion, in general, however, it is a function of time and one can write this as $\mathbf{v}(t)$.

Taking equations (3-2) and (3-4) together, and making use of the definition of *differentiation*, one finds that the velocity is nothing but the *derivative* of the vector function $\mathbf{r}(t)$ (see sec. 2.11.1):

$$\mathbf{v} = \frac{d\mathbf{r}}{dt}, \quad (3-6)$$

and signifies the instantaneous rate of change of the position vector of the particle.

Making use of equation (3-1) in this expression, one can express the velocity in terms of the rates of change of the co-ordinates x , y , and z of the particle:

$$\begin{aligned} \mathbf{v} &= \frac{dx}{dt}\hat{i} + \frac{dy}{dt}\hat{j} + \frac{dz}{dt}\hat{k} \\ &= v_x\hat{i} + v_y\hat{j} + v_z\hat{k}, \end{aligned} \quad (3-7)$$

where $v_x(= \frac{dx}{dt})$, $v_y(= \frac{dy}{dt})$, and $v_z(= \frac{dz}{dt})$, the three components of velocity, are the rates of displacement along the three co-ordinate axes.

3.2.2.4 Speed

Notice in fig. 3-3 that the magnitude of displacement in time T , i.e., the length of the segment PQ, is not the same as the distance covered by the particle along the arc PQ of its trajectory in the same time (for a uniformly moving particle, however, the two

are equal since the trajectory is a straight line then). Denoting the distance traversed along the trajectory by l , the average rate of traversal along the trajectory is seen to be $\frac{l}{T}$. If now one goes over to the limit $T \rightarrow 0$ as before, then the limiting value gives the instantaneous rate of traversal at time t . One can define this as the *speed* (say, u) of the particle:

$$u = \lim_{T \rightarrow 0} \frac{l}{T}. \quad (3-8)$$

Noting that, in the limit $T \rightarrow 0$, the chord PQ differs little from the arc along the trajectory, i.e., $l \rightarrow |s|$, and making use of equations (3-5), (3-8), one concludes that the speed, as defined above, is simply the magnitude of the velocity:

$$u = |\mathbf{v}| = v. \quad (3-9)$$

Analogous to the fact that the velocity is more fully expressed as $\mathbf{v}(t)$ to signify its time dependence, the speed may also be written as $v(t)$ to indicate that it is, in general, a function of time.

If the particle under consideration moves along its trajectory at a uniform rate then its speed, and hence the magnitude of its velocity, will remain unchanged with time but the *direction* of motion may, in general, keep on changing. Thus, even if the speed (v) remains constant, the velocity (\mathbf{v}) may change with time. On the other hand, motion with a constant *velocity* necessarily implies a constant speed since in this case the magnitude as well as the direction has to remain unchanged.

3.2.2.5 Planar and rectilinear motions

Referring to a set of co-ordinate axes OX, OY, OZ in any given frame of reference, the displacement of a particle implies, in general, a displacement along each of the three axes. If, however, the motion in the given frame happens to be such that the trajectory remains confined to a plane or a straight line then that motion is referred to as a *planar*

or a *rectilinear* motion respectively.

1. In general, the trajectory of a particle under a given force (refer to sec. 3.5.1) is a *space curve*. For some particular initial condition, or one belonging to some class of initial conditions, the space curve may be confined to a plane or to a straight line, but this cannot, in general, be so for *all* initial conditions. Or, there may be forces and initial conditions for which the space curve remains confined to some two dimensional surface (such as the surface of a sphere) or some curve of relatively simple description lying on a surface (such as a circle in a plane). These are instances of a more general type compared to planar or rectilinear motion, and are referred to as *two dimensional* motion and *one dimensional* motion respectively.
2. On the other hand, there exist special types of *fields of force* (refer to mech-field-sec), under which the trajectory remains confined to some plane or other for *all* initial conditions, where different classes of initial conditions correspond to different planes of motion.

If the motion be a planar one then, among the various possible co-ordinate systems in the frame of reference under consideration, one can choose a particular co-ordinate system for which the plane of motion remains confined to the x-y plane of the chosen frame. Similarly, for rectilinear motion the axes can be so chosen that the trajectory lies along the x-axis. Hence, the position vector $\mathbf{r}(t)$ can be written as, respectively,

$$\mathbf{r}(t) = x(t)\hat{i} + y(t)\hat{j}, \quad (3-10a)$$

$$\mathbf{r}(t) = x(t)\hat{i}. \quad (3-10b)$$

In the case of rectilinear motion, the quantity $x(t)$, which carries its own sign, completely determines the position vector $\mathbf{r}(t)$, and its derivative $\frac{dx}{dt}$, which also carries its own sign, completely determines the velocity \mathbf{v} . One often denotes this with the symbol v ($v \equiv \frac{dx}{dt}$), but *this* v differs from the speed in equation (3-9) which is now to be written as $|v|$ ($= |\frac{dx}{dt}|$). The one dimensional velocity v is positive or negative according as the direction

of motion of the particle is towards the positive or the negative direction of the x-axis (commonly, the positive direction is chosen by convention). In the special case of a particle moving uniformly along the x-axis, if x_1 and x_2 be its co-ordinates at times t_1 and t_2 respectively, then one can write

$$v = \frac{x_2 - x_1}{t_2 - t_1}. \quad (3-11)$$

One can make use of the instantaneous position co-ordinates to go through all relevant calculations for two dimensional and three dimensional motions as well. However, the use of the position vector and its derivatives is often found to be of considerable convenience.

Problem 3-2

Two particles, A and B, are located at time $t = 0$ at the origin of a co-ordinate system, and subsequently move along the x-y plane with velocities 1.0 and 2.0 (all quantities in SI units) along lines making angles $\frac{\pi}{6}$ and $\frac{\pi}{3}$, respectively, with the x-axis. Find the position vector of A with reference to B at $t = 5.0$.

Answer to Problem 3-2

HINT: The velocity vectors of the two particles are, $\mathbf{v}_A = 1.0 \frac{\sqrt{3}}{2} \hat{i} + 1.0 \frac{1}{2} \hat{j}$, and $\mathbf{v}_B = 2.0 \frac{1}{2} \hat{i} + 2.0 \frac{\sqrt{3}}{2} \hat{j}$. The instantaneous position vectors at time t are $\mathbf{r}_A = \mathbf{v}_A t$, $\mathbf{r}_B = \mathbf{v}_B t$. The required relative position vector is then $\mathbf{r}_{AB} = 5.0(\mathbf{v}_A - \mathbf{v}_B)$.

3.2.2.6 Momentum

The *mass* of a particle is of especial relevance in the description and explanation of its motion. Mass is a measure of *inertia*. More precisely, the *acceleration* of the particle under a given *force* (see sections 3.2.2.8 and 3.5.1 for these concepts) is determined by its mass.

If m denotes the mass of the particle, then its *momentum* at any instant of time is

defined by the expression $m\mathbf{v}$, where \mathbf{v} stands for its velocity at that instant. Denoting the momentum by the symbol \mathbf{p} , one has

$$\mathbf{p} = m\mathbf{v}. \quad (3-12)$$

This means that the Cartesian components of momentum are

$$p_x = mv_x = m \frac{dx}{dt}, \quad p_y = mv_y = m \frac{dy}{dt}, \quad p_z = mv_z = m \frac{dz}{dt}. \quad (3-13)$$

In the case of rectilinear motion, the momentum is completely specified by $p \equiv m \frac{dx}{dt}$, which carries its own sign, and differs from $|\mathbf{p}|$, just as $v \equiv \frac{dx}{dt}$ differs from the speed. The meanings carried by the symbols v and p will, in general, be clear from the context.

Notice that the momentum is not an independent kinematic quantity since it is completely determined by the mass m and velocity \mathbf{v} . However, it carries a special significance in mechanics and its use is often found to be more convenient as compared to that of the velocity \mathbf{v} .

3.2.2.7 Kinetic energy

One other physical quantity of importance in mechanics is the *kinetic energy* which, for a single particle, is defined as

$$K = \frac{1}{2}mv^2 = \frac{1}{2}m\mathbf{v} \cdot \mathbf{v}, \quad (3-14)$$

where $v(=|\mathbf{v}|)$ is the speed. Expressed in terms of momentum, the kinetic energy of a particle appears as

$$K = \frac{p^2}{2m} = \frac{1}{2m}\mathbf{p} \cdot \mathbf{p}, \quad (3-15)$$

where $p(=|\mathbf{p}|)$ is the magnitude of the momentum.

The significance of kinetic energy in mechanics will be indicated later in this chapter.

3.2.2.8 Acceleration

As I have mentioned above, the velocity of a particle can vary with time t . Denoting the velocities at time t and $t + T$ as $\mathbf{v}(t)$ and $\mathbf{v}(t + T)$ respectively, the *change* in velocity during the interval T is seen to be $\mathbf{v}(t + T) - \mathbf{v}(t)$, and so the average rate of change in this interval of time is $\frac{\mathbf{v}(t+T) - \mathbf{v}(t)}{T}$. If the interval T is imagined to be very small, or infinitesimal, then the above ratio gives the instantaneous rate of change of velocity at time t , and is termed the *acceleration* of the particle at that instant. Denoting this by the symbol \mathbf{a} , the definition of acceleration can be expressed in the form

$$\mathbf{a} = \lim_{T \rightarrow 0} \frac{\mathbf{v}(t + T) - \mathbf{v}(t)}{T}, \quad (3-16a)$$

i.e., making use of the definition of derivative of a vector (refer to sec. 2.11.1),

$$\mathbf{a} = \frac{d\mathbf{v}}{dt}. \quad (3-16b)$$

Expressed in terms of the derivatives of the velocity components the acceleration appears as

$$\mathbf{a} = \frac{dv_x}{dt} \hat{i} + \frac{dv_y}{dt} \hat{j} + \frac{dv_z}{dt} \hat{k}. \quad (3-17)$$

In the special case of rectilinear motion along, say, the x-axis, the acceleration is specified completely by

$$a = \frac{dv}{dt}, \quad (3-18)$$

i.e., the time derivative of the velocity v , as defined above for one dimensional motion. In this equation, a and v carry their own signs and differ from the magnitudes of the respective vectors, and the signed scalar a may be termed the acceleration in the same sense as v and p have been introduced above as the velocity and momentum for the one dimensional motion.

Analogous to the position vector \mathbf{r} and velocity \mathbf{v} , the acceleration \mathbf{a} also depends, in

general, on time t , and this time dependence can be expressed by writing it as $a(t)$.

However, there may be special situations where the acceleration does *not* change with time. The particle is then said to move with *uniform acceleration*. In the case of motion with uniform acceleration, one need not go over to the limit $T \rightarrow 0$ in equation (3-16a), and any finite value of T may be chosen. If, moreover, the motion be a rectilinear one, then one can write in (3-18),

$$a = \frac{v_2 - v_1}{t_2 - t_1}, \quad (3-19)$$

where v_1 and v_2 stand for the velocities (in the sense indicated above) at time instants t_1 and t_2 respectively.

Problem 3-3

A particle moving along the x-axis has the following position co-ordinates, determined at intervals of 0.1 s starting at $t = 0$ up to $t = 0.5$ s: x (in 10^{-3} m) = 0, 1.0, 2.1, 2.9, 3.8, 4.9. Obtain the average velocity in this interval of time and estimate the velocity and acceleration at $t = 0.3$ s.

Answer to Problem 3-3

An estimate for the average velocity in the interval between 0 and 0.5 s is $v_{av} = \frac{4.9}{0.5} \times 10^{-3} \text{ m}\cdot\text{s}^{-1}$. The average velocity in the interval between 0.2s and 0.4s, calculated similarly, gives an estimate of the instantaneous velocity at 0.3 s (answer: $0.85 \text{ m}\cdot\text{s}^{-1}$). An estimate for the instantaneous acceleration at 0.3 s is obtained in a similar manner from the velocities at 0.2 s and 0.4 s (answer: $0.025 \text{ m}\cdot\text{s}^{-2}$).

3.3 Newton's laws of motion: Introduction

As I have mentioned, the description of motion of a particle or any other mechanical system differs in various different frames of reference. As a result, one has to clearly identify the chosen frame of reference while formulating the basic equations in terms

of which the motion of the system can be understood. It is worthwhile to approach Newton's *first* and *second* laws of motion from this perspective.

Among the innumerable different frames of reference in which one can possibly describe the motion of a particle, Newton's *first law* identifies one particular set of frames as being of special relevance. Choosing any one frame from this set, the *second law* gives the basic rule according to which the motion of the particle is determined in such a frame. These special frames of reference are termed *inertial frames*. Thus, the second law serves the purpose of formulating the basic equations relating to the motion of a particle in an inertial frame.

It is important to realize, however, that the first law is not *just* a definition of the inertial frames. If it were nothing more than a definition, it would not have been of much significance. In reality, if the earth is assumed to be a rigid body, and one chooses a frame of reference rigidly attached to the earth, then it is found that in *that* frame of reference, the motion of a particle is quite well described by Newton's second law, which I will come to presently. In other words, the earth-bound frame is, to a good approximation, an inertial frame though, in a strict sense, one finds *some* discrepancy between the second law and the observed motion of a particle in the earth-bound frame. Instead, if one thinks of a frame fixed or uniformly moving with respect to the distant stars, then in that frame the discrepancies are found to be much smaller. The first law serves the purpose of thus identifying a set of frames that can *in practice* be considered to be inertial to a sufficiently good degree of approximation. In this book, we shall interpret the first law from this point of view.

3.4 Inertial frames. Newton's first law.

It is now time to underline the defining characteristic of an inertial frame. For this, one needs the concept of a *free* particle. Imagine a particle away from the influence of all other particles and systems in the universe, i.e., one that has no interaction with other systems around it. Such a particle will be termed a *free* one. Evidently, the concept of a free particle is an idealization. However, a particle located at a sufficiently large distance

from all other systems, can effectively be assumed to be a free particle.

An inertial frame of reference is then one in which a free particle either remains at rest or moves with uniform velocity.

A particle at rest is simply a special instance of motion with uniform velocity. We have already seen that a particle with uniform velocity moves along a straight line. If the velocity along that line is denoted as v , then a stationary particle corresponds to the special case $v = 0$. In the following, the particle at rest in any given frame will have to be understood as included when I speak of one in uniform motion.

With this definition of an inertial frame in mind, the following statement will, in this book, be taken to constitute the principal content of *Newton's first law of motion*:

A frame of reference in uniform motion with respect to distant stars is, to a good degree of approximation, an inertial frame.

More commonly, however, the first law is stated in a different form. In such a statement a frame uniformly moving with respect to the distant stars (or, to a slightly less accurate approximation, an earth-bound frame) is tacitly assumed to be the frame of reference under consideration. One then speaks of a free particle in this frame, meaning thereby that the particle is not subjected to any 'force', i.e., in other words, that its motion is not affected by the influence of other bodies. The first law then states that such a particle moves uniformly in this chosen frame. This actually implies that the frame chosen is an inertial one.

3.5 Newton's second law: Equation of motion

3.5.1 The concept of force

Since a free particle moves uniformly in an inertial frame, its *acceleration* in such a frame is zero. If, then, one finds that the acceleration of a particle in an inertial frame is

not zero, what would one conclude? Evidently, the particle can no longer be a free one, and hence must be under the influence of other bodies. And it must be this influence of other bodies that has to be held responsible for the non-zero acceleration of the particle. Such influence of other bodies on the particle under consideration is referred to as the *force* acting on it. This, of course, is not a quantitative definition of force, and is just a qualitative idea. The force acting on a particle at any given instant is, in reality, a vector quantity. Let us, for the moment, assume that there exists some well defined operational procedure for the quantitative determination of force on the particle, and let the force acting on the particle under consideration, determined in accordance with this procedure at any given instant of time be \mathbf{F} .

In a similar manner, let us also assume that there is known a well defined operational procedure for the determination of the mass m of the particle. We denote this mass as m and the acceleration of the particle at the chosen instant of time as \mathbf{a} .

3.5.2 The second law

Newton's second law of motion can then be expressed in the following mathematical form:

$$m\mathbf{a} = \mathbf{F}. \quad (3-20)$$

It states that the acceleration of a particle produced by the influence of other bodies on it, is proportional to the force acting on it, and the acceleration is directed along the line of action of the force. Here the constant of proportionality between force and acceleration is the mass of the particle. I remind you that this statement presupposes that the frame of reference in which the motion of the particle is being described, is to be an inertial one, where the *first law* identifies for us a set of inertial frames.

The second law is commonly stated in a slightly different form which is, in a sense, of more general scope: the rate of change of *momentum* is proportional to the force acting on the particle, and takes place in the direction in which the force acts.

Following our approach of defining the instantaneous velocity and acceleration, we consider the average rate of change of momentum of the particle during an interval from, say t to $t + T$ given by $\frac{\mathbf{p}(t+T) - \mathbf{p}(t)}{T}$, where $\mathbf{p}(t)$ and $\mathbf{p}(t + T)$ denote the momenta at times t and $t + T$ respectively. If now the interval T is imagined to be infinitesimally small, then this ratio reduces to the instantaneous rate of change of momentum at time t :

$$\frac{d\mathbf{p}}{dt} = \lim_{T \rightarrow 0} \frac{\mathbf{p}(t + T) - \mathbf{p}(t)}{T}. \quad (3-21)$$

This, of course, agrees with the definition of derivative of a vector stated in sec. 2.11.1.

3.5.3 Equation of motion

With this definition of the rate of change of momentum, the alternative form of equation (3-20), corresponding to the second law as stated commonly, is seen to be

$$\frac{d\mathbf{p}}{dt} = \mathbf{F}. \quad (3-22)$$

In this formula, there might have been present a constant of proportionality relating the force to the rate of change of momentum. However, that constant may be chosen to be unity, with an appropriate choice of the unit of force. In the SI system, this unit of force is the *newton* (N). Assuming that the unit of mass in the SI system has been defined beforehand (this being the kilogram (kg), which represents some definite quantity of matter), the force required to produce an acceleration of $1 \text{ m} \cdot \text{s}^{-2}$ in a body of unit mass is defined to be of magnitude 1 N (refer to equation (3-20)). The equation (3-22) is referred to as the *equation of motion* of the particle under consideration.

Newton's second law is sometimes stated with reference to the motion of a *body* instead of that of a particle. It is more logical, however, to start from a particle and then to go over to the motion of a body, where the latter is looked upon as a system made up of a number of particles. This involves a number of conceptual and mathematical steps, some of which we will briefly come across in a later section. In the end, an equation analogous to (3-22) is obtained, describing the motion of what is referred to as the

center of mass of the body. This can be taken as a partial description of the motion of the body as a whole since the motion of the center of mass does not describe the motion of the body completely. For instance, in the case of a *rigid* body, there may occur, in addition to the motion of the center of mass, a *rotation* around the center of mass. When one speaks of formula (3-22) being applied to a *body*, one actually means the motion of the center of mass. Of course, (3-22) remains applicable to each individual particle making up the body.

For a particle of constant mass, equations (3-20) and (3-22) are of identical content, since one has

$$\frac{d\mathbf{p}}{dt} = \frac{d}{dt}(m\mathbf{v}) = m\frac{d\mathbf{v}}{dt} = m\mathbf{a}, \quad (3-23)$$

where equation (3-16b) has been made use of. Unless otherwise stated, we shall assume in this book that equations (3-20) and (3-22) can be used interchangeably.

1. In the *relativistic* theory, however, the mass is found to be dependent on the velocity of the particle, as a result of which the second equality in (3-23) is no longer valid, and (3-22) can no longer be taken to be equivalent to (3-20). One then takes (3-22) as the more generally valid form of Newton's second law. I should, however, mention here that the relativistic theory has a different setting as compared to the non-relativistic one, and the definition of relevant physical quantities as also the relations between these quantities in the relativistic setting involve expressions of a more general mathematical structure (refer to chapter 18).
2. In the non-relativistic theory one encounters, at times, a body or a system of *variable* mass like, for instance, a snowball gradually gaining in size. In describing the motion of the *center of mass* of such a body one obtains an equation of the form (3-22) where, once again the second equality of (3-23) cannot be invoked.

In summary, while the formula (3-22) is to be looked at as the more appropriate form of the second law, we shall be concerned, in this book, mostly with non-relativistic motion of particles or bodies of constant mass where the above equivalence of (3-20) and (3-22) can be assumed to hold.

The reason why equation (3-22) is referred to as the equation of motion of the particle is that it allows one to *infer* the state of motion of the particle at any arbitrarily chosen time t , once its state of motion is known at a given instant, say t_0 , where t can be either prior to or later than t_0 . This is what can be looked at as a *causal* explanation of the motion of the particle. If the force acting on the particle be looked upon as a *cause* and if this ‘cause’ is known in a given context, then the state of motion at any chosen time can be interpreted as an *effect* that can be inferred from the second law once a set of *initial conditions* corresponding to the state of motion at some given time t_0 are known.

In this context, the term *state of motion* at any given instant refers to the position vector (\mathbf{r}) and the velocity (\mathbf{v}) of the particle, taken together, at that instant. At times, it is more convenient to use the momentum (\mathbf{p}) in place of the velocity. Choosing any appropriate co-ordinate system in the frame of reference under consideration, the state of motion is specified in terms of the three components of \mathbf{r} , along with the three components of \mathbf{p} or \mathbf{v} .

Equation (3-22) is an equality between vector quantities and can alternatively be written as three scalar equations relating the components of these vector quantities. Thus,

$$ma_x = F_x, \quad ma_y = F_y, \quad ma_z = F_z, \quad (3-24)$$

where a_x, a_y, a_z and F_x, F_y, F_z stands for the Cartesian components of the acceleration and force respectively, relative to any chosen co-ordinate system. In terms of the components of the momentum, one has

$$\frac{dp_x}{dt} = F_x, \quad \frac{dp_y}{dt} = F_y, \quad \frac{dp_z}{dt} = F_z. \quad (3-25)$$

3.5.4 The line of action of a force

In introducing the concept of force and setting up the equation of motion of a particle, we noted that the force is a vector quantity, i.e., in order to describe a force, one has to specify its direction as well as its magnitude.

In addition, one needs to specify the *unit* so as to describe a force as a physical quantity. In the SI system the unit of force is the *newton* (N).

However, this still does not give a complete description of a force, because a force is characterized by some definite *line of action* as well. Fig 3-4 depicts two forces with a pair of directed line segments, both having the same direction and magnitude. Thus, any one of the segments can be made to coincide with the other by a parallel translation. In this instance, the directed line segments represent the same vector, but *not* the same force, because their lines of action are different. In other words, though the *vectors* \mathbf{F}_1 and \mathbf{F}_2 are equal, the *forces* \mathbf{F}_1 and \mathbf{F}_2 are distinct.

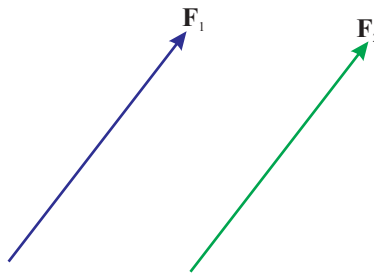


Figure 3-4: Two forces, \mathbf{F}_1 and \mathbf{F}_2 , having the same magnitude and direction, but different lines of action.

Any force arises by virtue of the action of a particle or a system of particles on another particle or a system of particles. We consider, for the time being, forces acting on single particles or on rigid bodies. If, in a system of particles, the constituent particles remain at fixed distances from one another, then it is termed a rigid body. In mechanics, the concept of a rigid body is an idealized but convenient one.

With reference to a force acting on a particle, the latter is commonly referred to as the *point of application* of the force. However, for a force applied on a particle or a rigid body, the point of application does not have any special relevance - once the line of action of the force is specified, any particle of the body lying on that line of action, or even any other point lying on the line of action and imagined to be rigidly attached to the particle or the body, may be taken as the point of application of the force. For the

special case of a force acting on a particle, however, we shall continue to refer to that particle itself as the point of application of the force. Specifying the point of application of a force in addition to its line of action assumes relevance in the definition of *work*.

3.5.5 The resultant of a number of forces

In mechanics, a force on a particle is known by the effect it produces on the state of motion of the particle or, in other words, by the resulting acceleration of the particle. Suppose that two forces \mathbf{F}_1 and \mathbf{F}_2 acting on a particle produce the same acceleration in it as that produced by a single force, say, \mathbf{F} acting on the same particle. Then \mathbf{F} is referred to as the *resultant* of the forces \mathbf{F}_1 and \mathbf{F}_2 . This requires that the force \mathbf{F} has to be the *vector sum* of the two forces \mathbf{F}_1 and \mathbf{F}_2 :

$$\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2. \quad (3-26)$$

1. This derives from the fact that the result of two simultaneous displacements is a single displacement given by the vector sum of the two (check this out). Since acceleration is the second derivative of the displacement with respect to time, the vector summation rule applies to acceleration as well. Finally, the equation of motion tells us that the force is related linearly to the acceleration, which explains why the vector sum formula gives the resultant of two forces acting on a particle.
2. The two forces \mathbf{F}_1 and \mathbf{F}_2 on a given particle, say, P , may arise due to the action of two other particles, say Q and R , on it. A basic assumption in mechanics is the *independence* of the two forces exerted by Q and R .

More generally, if a number of forces, say, $\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_N$ act on a particle, being exerted by N number of other particles or systems of particles, then the effect of these on the particle under consideration is the same in all respects as that of a single force

$$\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2 + \dots + \mathbf{F}_N, \quad (3-27)$$

acting on it. This gives the resultant of a number of forces, all having the same point of application or, more precisely, the lines of action of all being concurrent. Here the line of action of the resultant also passes through the common point of application of the forces (or, more precisely, the common point on their lines of action). The process of arriving at the resultant of a number of given forces is referred to as one of *composition* of these forces.

3.5.6 Forces in equilibrium

Suppose that two forces, say, \mathbf{F}_1 and \mathbf{F}_2 acting simultaneously on a particle fail to produce any change of its state of motion. For instance, assuming that the particle is initially at rest, suppose that it continues to be at rest under the action of the two forces. We shall then say that the two forces \mathbf{F}_1 and \mathbf{F}_2 form a system *in equilibrium*. Evidently, the vector sum of the two forces has to be zero for these to be in equilibrium. In a similar manner, if a number of forces, say, $\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_N$, act simultaneously on a particle then the condition for the state of motion of the particle to remain unchanged in spite of the action of these, i.e., the condition for *equilibrium* of the system of forces is

$$\mathbf{F}_1 + \mathbf{F}_2 + \dots + \mathbf{F}_N = 0. \quad (3-28)$$

This is the basic equation relating to the equilibrium of a system of forces acting on a particle, i.e., for a system of concurrent forces. More general considerations are necessary to formulate the condition of equilibrium of a system of forces acting on a rigid body, where the forces need not be concurrent. I shall come back to the question of equilibrium of concurrent and non-concurrent forces later in section 3.22.

Problem 3-4

Two forces, $\mathbf{F}_1, \mathbf{F}_2$, acting on a particle have Cartesian components (measured in newton) 2.5, -3.0, 1.7, and -1.5, 2.5, 0.5, respectively. Find the components of a third force, \mathbf{F}_3 such that the three forces form a system in equilibrium.

Answer to Problem 3-4

ANSWER: $\mathbf{F}_3 = -1.0\hat{i} + 0.5\hat{j} - 2.2\hat{k}$ (in newton).

3.6 Motion along a straight line

The equations of motion appear in relatively simple forms for rectilinear and planar motions. We consider first the special case of one dimensional motion along a straight line. The condition for such a motion to take place is that the force on the particle at any and every instant of time during its motion should act along its direction of motion, determined at any chosen initial time. The motion will then be confined in the straight line drawn through the initial position along the direction of the initial velocity vector.

While we are considering here the special case of a rectilinear motion, one needs to be careful in using the term ‘one dimensional motion’ since it may mean different things in different contexts (refer to sec. 3.2.2.5).

Recall that, in general, for any given initial state of motion, the trajectory of the particle, obtained by solving the equations of motion, will be a curve in space, the motion on which can be described in terms of one single time dependent co-ordinate. However, this curve is not necessarily one confined to a plane, while rectilinear motion is even more of a special case. Still more rare is the case where the motion remains confined to a *fixed* planar curve or a fixed straight line regardless of the initial conditions, for which one or more *constraints* are to act on the particle, where a constraint is some kind of a restrictive condition operating on it.

In the present context, however, the term ‘motion along a straight line’ is used without reference to the above distinction. It may so happen, for instance, that a particle, acted upon by a given force, describes rectilinear trajectories parallel to a given direction in space for a given set of initial conditions while it describes various different space curves for other initial conditions. Much of the present discussion will apply to a description of such a particular class of rectilinear trajectories, while it also applies to a description of *constrained* motion along a straight line. In other words, we do not enter into the question of how in reality a particle may turn out to execute a

rectilinear motion along a given direction. We simply assume that the forces and initial conditions are such that the particle moves along a straight line parallel to a given direction, which one may take to be along the x-axis of a co-ordinate system.

As mentioned above, let a Cartesian co-ordinate system be chosen with its x-axis along the above straight line. Denoting then the components of displacement, velocity, momentum, acceleration, and force along this line as, respectively, x , v , p , a , F , the equation of motion assumes the simple form

$$\frac{dp}{dt} = F. \quad (3-29a)$$

or, assuming the mass of the particle to be constant,

$$ma = m \frac{dv}{dt} = m \frac{d^2x}{dt^2} = F, \quad (3-29b)$$

This is nothing but the first equation in (3-24) where the second and third equations are redundant due to the fact that the motion is one dimensional. Consequently, the task of determining the motion of the particle by solving the equation of motion becomes simpler.

In view of the common practice of denoting the magnitude of a vector by the same symbol as the vector itself, but written without using the boldface, I must once again remind you that the symbols like v , p , a , F for one dimensional motion do not stand for the magnitudes of corresponding vectors, but for their x-components. Stated differently, these are the magnitudes of the vectors along with appropriate *signs* that tell us whether the vectors point to the positive or the negative directions of the x-axis.

A special instance of a rectilinear motion is a motion *with uniform acceleration*. For such motion neither the force F nor the acceleration a changes with time and the velocity changes at a *uniform rate*. In this case, if v_1 and v_2 be the velocities at any two given time instants t_1 and t_2 respectively, then a is given by (3-19). Thus, for a given force F

or equivalently, for a given value of $a(= \frac{F}{m})$, one has

$$v_2 = v_1 + a(t_2 - t_1). \quad (3-30a)$$

Referring to t_1 as the *initial* time, v_1 is termed the initial velocity, and is sometimes denoted by the symbol u . Similarly, denoting the time interval $t_2 - t_1$ by t , v_2 is referred to as the *final* velocity after an interval t , being denoted by the symbol v , where the relation between u , v and t is

$$v = u + at. \quad (3-30b)$$

If, moreover, x_1 and x_2 be the position co-ordinates of the particle at the initial and final instants of the interval, then the displacement in this interval is seen to be

$$s = x_2 - x_1, \quad (3-31)$$

this being the displacement along the x-axis, where the displacements along the y- and z-axes are both zero.

Let us now determine the expression for the displacement in terms of the time interval t for uniformly accelerated motion along a straight line. Only in the special case of *uniform* motion ($a = 0$) with velocity, say, v is this relation obtained straightaway as the product of velocity and time interval, i.e.,

$$s = vt. \quad (3-32)$$

For a non-zero but constant acceleration, the velocity keeps on changing with time but in a simple manner, namely, at a uniform rate. As a result, the displacement in a given time interval is obtained as the product of the *average* velocity and the time interval t . According to the formula (3-30b), the average velocity is given by

$$v_{\text{average}} = \frac{u + v}{2} = u + \frac{1}{2}at, \quad (3-33)$$

and so the displacement works out to

$$s = ut + \frac{1}{2}at^2. \quad (3-34)$$

Together, formulae (3-30b) and (3-34) give a complete description of the rectilinear motion of a single particle under a constant force F , corresponding to which the acceleration is given by $a = \frac{F}{m}$.

Here is one more useful relation that holds for rectilinear motion with uniform acceleration,

$$v^2 = u^2 + 2as, \quad (3-35)$$

where the meanings of the symbols are as above (check this relation out).

Problem 3-5

A particle moves along the x-axis of a co-ordinate system, starting from rest at $x = 4.0\text{m}$ and at time $t = 0\text{s}$, with an acceleration of $1.0\text{m}\cdot\text{s}^{-2}$. A second particle starts from rest at $x = 0$ and at time $t = 2\text{s}$, with an acceleration of $2.0\text{ m}\cdot\text{s}^{-2}$. If the accelerations be uniform, at what time will the second particle catch up with the first?

Answer to Problem 3-5

HINT: If the two particles meet at time $T\text{s}$ after the second particle starts moving, then $\frac{1}{2} \times 2 \times T^2 = 4.0 + \frac{1}{2} \times 1 \times (T + 2)^2$. The positive root for T of this quadratic equation works out to $T = 6$.

Problem 3-6

Two particles, A, B, are at a distance s apart at time $t = 0$. A moves towards B with a uniform velocity v , while B moves away from A with an initial velocity u and a uniform acceleration a , where $u < v$. What is the minimum value of a such that A does not catch up with B. Assume that both particles move along a given straight line.

Answer to Problem 3-6

HINT: In order that A may catch up with B, the relation $vt = s + ut + \frac{1}{2}at^2$ is to be satisfied for some positive real value of t . Thus, A will never catch up with B if $a > \frac{(v-u)^2}{2s}$. For a less than $\frac{(v-u)^2}{2s}$, there are two positive solutions for t , signifying that after A meets B and overtakes it (for this, B has to be transparent!), the two particles again meet by virtue of the acceleration of the latter.

A motion with a uniform acceleration is, evidently, a special one. More generally, the force need not remain constant in a rectilinear motion. If $F(x)$ denotes the force at position x , then the equation of motion, eq. (3-29b), can be written in the form

$$m \frac{d^2x}{dt^2} = F(x). \quad (3-36a)$$

One has to solve this differential equation by performing two successive integrations so as to obtain the position and velocity of the particle at any instant of time t in terms of the position and velocity at a chosen initial instant $t = 0$ (or, say, $t = t_0$).

One such instance of rectilinear motion under a non-constant force is met with in *simple harmonic motion*, which we shall look at in chapter 4. In numerous situations of interest, the force depends explicitly on time as well, when the equation of motion assumes the form

$$m \frac{d^2x}{dt^2} = F(x, t). \quad (3-36b)$$

An instance of a one dimensional motion under a time dependent force will be found in sec. 4.6 in connection with the forced simple harmonic motion where, moreover, the force on the particle at any instant will be seen to depend on the *velocity* of the particle as well.

The following problem will give you an idea as to what it means to solve the equation of one dimensional motion by performing two successive integrations so to obtain the position and velocity of the particle at any given instant of time.

Problem 3-7

Consider a particle of mass m moving along the x -axis under a time dependent force given by $F(t) = at$, where a is a given constant. The particle starts from rest at $x = x_0$ at time $t = 0$. Find its velocity and position at any instant of time t .

Answer to Problem 3-7

HINT: The equation of motion under the given force is $m \frac{d^2x}{dt^2} = at$. Integrating with respect to time, we get $m \frac{dx}{dt} = \frac{at^2}{2} + A$, where A is a constant of integration. Making use of the initial condition that the velocity is zero at $t = 0$, one obtains $A = 0$. This gives the velocity at time t as $v = \frac{dx}{dt} = \frac{at^2}{2m}$. Integrating once again with respect to time, the position at time t is seen to be $x = \frac{at^3}{6m} + B$, where B is a second constant of integration. Making use of the initial condition that $x = x_0$ at $t = 0$, one finds $B = x_0$. Thus, the position of the particle at time t is given by $x = x_0 + \frac{at^3}{6m}$.

3.7 Motion in a plane

If the motion of the particle is confined to a plane, it is said to be a planar motion, which is a special instance of motion in two dimensions.

As I in the case of motion along a straight line, there may or may not be a constraint involved in the case of motion in a plane. Whenever the space curve describing the trajectory of the particle remains confined to a plane, which need not be a *fixed* plane independent of the initial conditions, it will be considered as an instance of planar motion (depending on the context, however, the term planar motion may be used in a different sense).

As an example, a particle moving under the gravitational attraction of a second, *fixed* particle stays confined to a plane that depends on its initial conditions, i.e., the initial state of motion. This is an instance of planar motion without constraint. On the other hand, a particle made to move on a table-top provides an instance of constrained planar motion since here the plane is a fixed one (i.e., is independent of the initial conditions on this plane) and all admissible initial conditions give rise to trajectories lying in this plane. From a broad point of view, however, the two are not different since

a constrained motion is similar to an unconstrained one with the added feature that here a special type of force, the *force of constraint*, operates on the particle.

For our present purpose, a planar motion is one for which the trajectory remains confined to a plane that may depend on the initial conditions.

Problem 3-8

The instantaneous position vector of a particle in a Cartesian co-ordinate system is $\mathbf{r}(t) = a \cos \phi \cos \omega t \hat{i} + a \sin \omega t \hat{j} - a \sin \phi \sin \omega t \hat{k}$, where a, ω are constants. Show that the motion is a planar one, and find the orbit of the particle.

Answer to Problem 3-8

HINT: The constant vector $\hat{n} = \sin \phi \hat{i} + \cos \phi \hat{k}$ is perpendicular to $\mathbf{r}(t)$ at all times. This shows that the motion is confined to the plane perpendicular to \hat{n} . Consider a new co-ordinate system obtained by a rotation about the y-axis of the old system by an angle ϕ . Arguing in a manner similar to problem 2-11, the transformation matrix giving the co-ordinates in the new system in terms of those in the old one is

$$T = \begin{pmatrix} \cos \phi & 0 & -\sin \phi \\ 0 & 1 & 0 \\ \sin \phi & 0 & \cos \phi \end{pmatrix}. \quad (3-37)$$

In terms of these new co-ordinates, one finds $\mathbf{r}(t) = a \cos \omega t \hat{i}' + a \sin \omega t \hat{j}'$, which shows that the orbit is a circle ($x'^2 + y'^2 = a^2$) in a plane obtained by rotating the old x-y plane by an angle ϕ about the y-axis. The direction of motion on this circle is in the sense of a rotation from the x' -axis to the y' -axis (check this out by working out the velocity at $t = 0$); see fig. 3-5.

Suppose that the force on a particle is such that it is always parallel to a certain plane and that the initial velocity (i.e., the velocity at a given initial point of time) is also parallel to this plane. The trajectory of the particle will then remain confined to a plane parallel to the one mentioned above, which passes through the initial position of the particle.

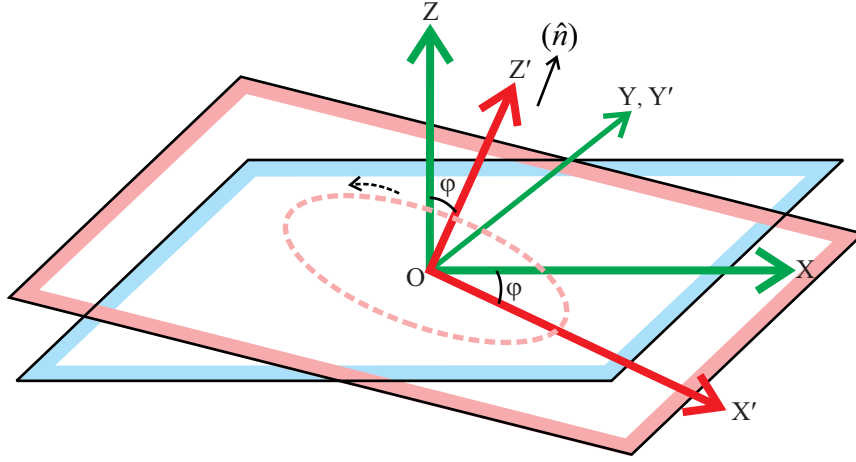


Figure 3-5: Illustrating the orbit of the particle for problem 3-8; the co-ordinate system with axes OX' , OY' , OZ' is obtained by a rotation of ϕ about the y -axis belonging to the system with axes OX , OY , OZ ; the x - y plane and the x' - y' planes are depicted; the orbit is a circle in the latter plane, the direction of motion on the circle being in the sense of rotation from the x' - to the y' -axis.

One can choose the co-ordinate system such that this particular plane coincides with the x - y plane of this system (see fig. 3-6). In that case the position vector (\mathbf{r}), velocity (\mathbf{v}), momentum (\mathbf{p}), acceleration (\mathbf{a}), and the force (\mathbf{F}) acting on the particle will all be two-dimensional vectors, the z -component of each being zero (for instance, $\mathbf{v} = v_x \hat{i} + v_y \hat{j}$). Consequently, one needs only two of the equations (3-24) (or (3-25)) for a complete description of the motion of the particle:

$$\frac{dp_x}{dt} = m \frac{dv_x}{dt} = F_x, \quad \frac{dp_y}{dt} = m \frac{dv_y}{dt} = F_y. \quad (3-38)$$

Here F_x , F_y may, in general, depend on the position co-ordinates of the particle in the plane.

As a special case of a planar motion one may mention the motion under a *constant* force where the initial velocity need not be parallel to the direction of that force (recall from section 3.6 that, if the initial velocity happens to be parallel to the force, then the motion reduces to a rectilinear one). Here F_x and F_y are constants independent of the position co-ordinates x , y , and the acceleration $\mathbf{a} = \frac{1}{m} \mathbf{F}$ is also a constant vector, with components a_x , a_y . In other words, the particle performs a two dimensional motion

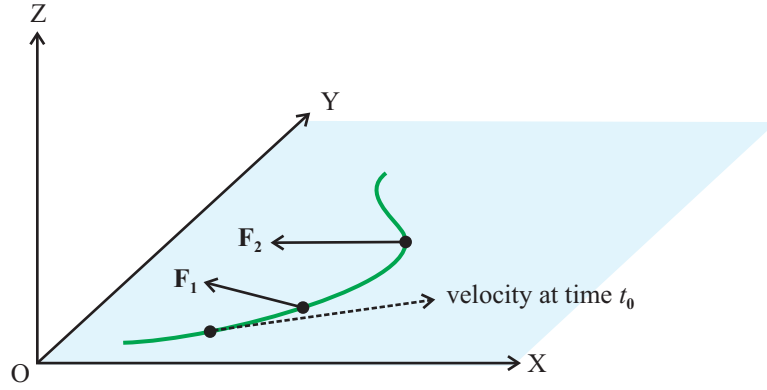


Figure 3-6: Two dimensional planar motion; the force on the particle is always parallel to the plane through the initial position, containing the initial velocity (the x-y plane here) where any arbitrarily chosen time t_0 can be taken as the initial time; F_1 and F_2 depict the force at two positions of the particle; the trajectory remains confined to the x-y plane.

with *uniform acceleration*. In this case, each of the two equations in (3-38) is seen to be independently similar to (3-29b), which means that the motion of the particle can be described as the resultant of two independent motions, one along each axis, each of the two being a rectilinear motion with uniform acceleration.

We adopt the following notation. Let u_0, v_0 denote the components of velocity at any chosen instant of time t_0 which we refer to as the *initial* time, it being assumed that the state of motion at this instant (alternatively referred to as the initial condition) is known. The velocity components at any other instant $t_0 + t$ are denoted as u, v (note the change in notation as compared to that in the case of one dimensional motion, and the changed significance of the symbols u, v as compared to those in sec. 3.6). We refer to $t_0 + t$ as the *final* time where this final time may, however, be *earlier* as well as later than the initial time, the time *interval* between the two being t . Finally, let x, y denote the displacements along the x- and y-axes respectively in the interval t .

Then, analogous to (3-30b) and (3-34), one has, for a planar motion under a constant force,

$$u = u_0 + a_x t, \quad v = v_0 + a_y t, \quad (3-39a)$$

$$x = u_0t + \frac{1}{2}a_x t^2, \quad y = v_0t + \frac{1}{2}a_y t^2, \quad (3-39b)$$

where x, y stand for the components of the displacement in interval t .

An instance of planar motion with uniform acceleration is met with in the motion of a *projectile* - a particle moving in a region sufficiently close to the earth's surface, the only force on it being the gravitational attraction of the earth. The acceleration due to this force is a constant, independent of the mass of the particle, and is directed vertically downward. It is referred to as the *acceleration due to gravity* (see chapter 5) and is commonly denoted by the symbol g , its value being approximately $9.8 \text{ m}\cdot\text{s}^{-2}$.

Problem 3-9

The motion of a projectile.

Show that the projectile undergoes a planar motion. It is customary to choose the plane of motion as the x-z plane, with the x-axis chosen horizontal and the z-axis along the vertically upward direction. Using a notation analogous to the one introduced above, show that final velocity components and the displacement components at the end of a time interval t are given by

$$u = u_0, \quad v = v_0 - gt \quad (3-40a)$$

$$x = u_0t, \quad z = v_0t - \frac{1}{2}gt^2. \quad (3-40b)$$

Show that this corresponds to a *parabolic trajectory* in the x-z plane..

Answer to Problem 3-9

The accelerations along the x- and z-directions are respectively zero and $-g$. The equation of the trajectory in the x-z plane is obtained by eliminating t in the equations (3-40b).

3.8 Field of force

In the last two sections we have seen that, starting from the equations of motion of the particle, one can arrive quite easily at a complete description of the motion, provided that the force on the particle is constant, independent of its position. More generally, however, the force acting on the particle does depend on the position, i.e., each of the components (F_x , F_y , F_z) of the force is a function of the three position co-ordinates x , y , z . Put another way, \mathbf{F} is a vector function of \mathbf{r} , and can be written as $\mathbf{F}(\mathbf{r})$. Imagining a directed line segment to be associated with each point in some region of space, where the segment represents the force on the particle when located at that point, one has a collection of directed line segments representing the forces at various points throughout that region, and one refers to such collection as a *field of force*. A field of force is a special instance of a *vector field* introduced in sec. 2.13.

Figure 3-7 depicts schematically a field of force where the arrows represent the forces at the respective points in a given region.

In characterizing a field of force, one has to distinguish between *conservative* and *non-conservative* fields. However, in order to make the distinction clear, one has to explain what is meant by *work* done by a force. I shall come back to this in sec. 3.11.

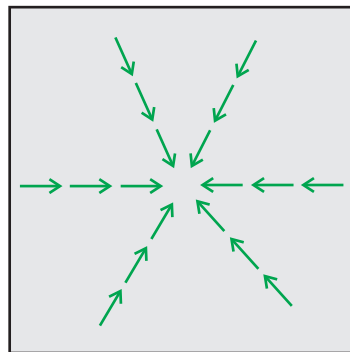


Figure 3-7: Schematic illustration of a field of force; the arrow at each point (with position vector, say, \mathbf{r}) in a specified region represents the force (say, \mathbf{F}) at that point; the field of force is mathematically represented as a vector function $\mathbf{F}(\mathbf{r})$.

3.9 Transformations from one frame of reference to another

I have earlier addressed the question of frames of reference and their significance in mechanics. While introducing Newton's second law and the equation of motion of a particle, I have mentioned the special role of a certain class of frames of reference termed the inertial frames. However, *non-inertial* frames are also relevant in mechanics in that the use of a non-inertial frame is, at times, convenient in understanding the motion of the particle or, more generally, of any dynamical system.

Suppose that S and S' are any two frames of reference, in each of which an appropriate co-ordinate system has been chosen, as in fig. 3-8. Evidently, the kinematic and dynamical quantities relating to the motion of a particle will differ in the descriptions of the motion from the two frames. However, since these two sets of quantities refer to the *same* physical phenomenon of the motion of the particle under consideration, they must necessarily be related to each other in some definite manner. The formulae expressing these relations are referred to as the *transformation* rules between the two frames.

3.9.1 Transformation of velocity

As an illustration of the transformation formulae, let us look at the instantaneous velocity of a particle as observed in the two frames. Let, at any given instant, \mathbf{v} and \mathbf{v}' be the velocities with reference to S and S' respectively, and let the velocity of S' relative to S at that instant be \mathbf{V} .

In defining \mathbf{V} , one needs to refer to some fixed point in S' , say, the origin O' of the co-ordinate system (see fig. 3-8) chosen in it. \mathbf{V} is then the instantaneous velocity of O' with respect to S . The relation between \mathbf{v} and \mathbf{v}' which constitutes the velocity transformation formula between S and S' will evidently depend on \mathbf{V} .

The relation between \mathbf{v} and \mathbf{v}' , commonly known as the *formula for relative velocities*, is

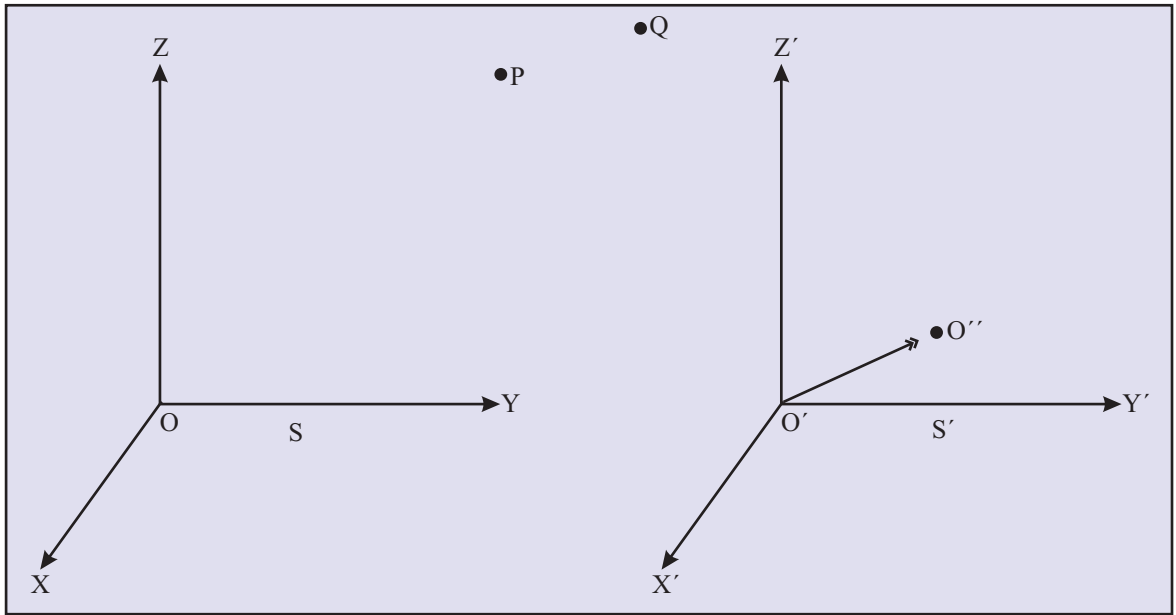


Figure 3-8: Illustrating the idea of transformation between two frames of reference S and S'; co-ordinate systems (OXYZ and O'X'Y'Z') have been chosen in the two frames; the double-headed arrow represents the translational velocity of S' relative to S, where it is assumed that there is no rotational motion; O'' denotes the position of O', as seen in S, at a later instant; $\vec{O'P}$ and $\vec{O'Q}$ are the position vectors of the particle relative to S at the two instants, while $\vec{O''P}$ and $\vec{O''Q}$ are the corresponding position vectors in S'; while we have considered the motion of S' relative to S, we could as well have considered the motion of S relative to S'.

not difficult to obtain, and can be written as

$$\mathbf{v}' = \mathbf{v} - \mathbf{V}. \quad (3-41)$$

Problem 3-10

Establish the formula(3-41).

Answer to Problem 3-10

HINT: Look at fig 3-8. If P and Q represent positions at an infinitesimally small interval δt , then $\vec{PQ} = \mathbf{v}\delta t$, while $\vec{O''Q} - \vec{O'P} = \mathbf{v}'\delta t$. Finally, $\vec{O'\vec{O''}} = \mathbf{V}\delta t$.

Written in terms of the components of the vectors with respect to the two co-ordinate

systems chosen, the above velocity transformation formula appears as

$$v'_x = v_x - V_x, \quad v'_y = v_y - V_y, \quad v'_z = v_z - V_z. \quad (3-42)$$

The relations (3-42) are based on the assumption that the respective axes of the two co-ordinate systems remain parallel to each other. In case these are not parallel, the relations (3-42) do not hold, though (3-41) continues to be valid.

However, even though the axes in S and S' need not be parallel for eq. (3-41) to hold, the validity of the latter requires that the *orientation* of the co-ordinate axes in S' must not change with time relative to the axes in S. For instance, suppose that the co-ordinate system fixed in S' is *rotating* about an axis fixed relative to S. One then says that S' is a rotating frame relative to S, and the formula (3-41) then needs modification. If S' does not possess such rotational motion relative to S at the instant under consideration, then the orientation of axes in S' does not change relative to those in S, and the transformation formula (3-41) for the velocities remains valid. In the absence of rotational motion, S' is said to possess only a *translational motion* relative to S. In general, however, the instantaneous motion of a frame S' relative to S can be a combination of a translation and a rotation.

3.9.2 Relative velocity

Let \mathbf{v}_1 and \mathbf{v}_2 be the instantaneous velocities of two particles (say P, Q) relative to any given frame of reference S. Considering now a frame of reference (say, S') in which P is instantaneously at rest, the velocity of Q in the frame S' is referred to as the *relative velocity* of Q with respect to P. Making use of eq. (3-41), the velocity of S' relative to the frame S is \mathbf{v}_1 in the present context, and thus the relative velocity of Q with respect to P is (again from eq. (3-41)); note, however, the change in the meanings of the symbols)

$$\mathbf{V} = \mathbf{v}_2 - \mathbf{v}_1, \quad (3-43)$$

In other words, the relative velocity of any two particles is simply the difference of their

velocities, taken in appropriate order.

What is interesting to note about the relative velocity is that, under a change of frame of reference, the velocities of the two particles under consideration get transformed to changed values, but their relative velocity remains unchanged (check this out, making use of eq. (3-41)): *the relative velocity of two particles at any given instant of time is the same in all frames.*

More generally, the difference of *any two* velocities, whether at the same or at different instants of time, observed in any given frame of reference (say, S_1), remains unchanged when transformed to another frame (say, S_2) *moving uniformly* with respect to S_1 . This result holds, in particular, when S_1 and S_2 are two *inertial* frames of reference (see sec. 3.10.1) since inertial frames move *uniformly* with respect to one another.

Incidentally, the formula (3-43) for relative velocities can be paraphrased in a way that makes it easy to remember: *the velocity of Q relative to S is its velocity relative to P (i.e., S_1) plus the velocity of P (i.e., S_1) relative to S.*

Problem 3-11

Rowing a boat across a river.

A person rowing a boat starts from a point P on one bank of a river which flows along the x-axis of a co-ordinate system, where the y-axis points towards the opposite bank. The river flows with a stream velocity v while the velocity of the boat relative to the stream is u , making an angle θ with the x-axis. If the width of the river be a , find the time (T) taken by the boat to cross the river and the distance (D) traversed by it in the process. What are the minimum values of T and D as functions of θ ?

Answer to Problem 3-11

HINT: The velocity components of the boat relative to land are $v_x = v + u \cos \theta$, $v_y = u \sin \theta$. The time taken to reach the opposite bank is then $T = \frac{a}{u \sin \theta}$. In this time the boat moves downstream by a distance $d = T(v + u \cos \theta) = \frac{a(v + u \cos \theta)}{u \sin \theta}$. The distance traversed is then $D = a \frac{\sqrt{(u^2 + v^2 + 2uv \cos \theta)}}{u \sin \theta}$.

When considered as functions of θ , the minimum values of T and D are respectively (a) $T_{\min} = \frac{a}{u}$ ($\theta = \frac{\pi}{2}$, rowing perpendicularly to the stream) and (b) $D_{\min} = a$ for $u > v$ ($\cos \theta = -\frac{v}{u}$, rowing upstream such that the velocity relative to the bank is along the y-axis), while $D_{\min} = a\frac{v}{u}$ for $u < v$ ($\cos \theta = -\frac{u}{v}$); in rowing for minimum distance, the time of travel may become very large, depending on the values of u and v .

NOTE: In the above solution, the determination of the minimum time, and also the determination of the minimum distance in the case $u > v$ follow from simple considerations relating to the geometry of the problem (try this out). The determination of the minimum distance in the case $u < v$, on the other hand, requires familiarity with elementary notions in the theory of maxima and minima of functions. In the present instance, the minimum value is to be determined by setting to zero the *derivative* with respect to θ .

Problem 3-12

Two particles, A and B, are located at points with position vectors $\mathbf{r}_1, \mathbf{r}_2$ at some instant of time, and move with uniform velocities $\mathbf{v}_1, \mathbf{v}_2$. Find the condition that they meet at some future time.

Answer to Problem 3-12

HINT: The relative velocity of B with reference to A is to be directed along the instantaneous position vector of A with reference to B, i.e., $\frac{\mathbf{r}_1 - \mathbf{r}_2}{|\mathbf{r}_1 - \mathbf{r}_2|} = -\frac{\mathbf{v}_2 - \mathbf{v}_1}{|\mathbf{v}_2 - \mathbf{v}_1|}$.

3.9.3 Transformations of displacement and time

As I have mentioned earlier, the measurement of time is independent of the frame of reference. Indeed, this is a fundamental assumption in non-relativistic mechanics. According to this assumption, the transformation formula for time as measured in two frames S and S' is

$$t' = t. \quad (3-44)$$

Here t and t' are the times, as measured in the two frames, corresponding to some particular *event*. For, instance, think of a particle being located at some particular point

in space at some particular instant of time. This is an event. The time corresponding to this event, as measured in S or S' , is then t or t' respectively. Here I have assumed that the times in the two clocks used in S and S' have been set so as to agree at some chosen instant, which is taken as $t = t' = 0$. This time setting is, however, not essential, because what is of ultimate relevance is the time *interval* between given events. If the interval between any two events be t and t' as measured by clocks in S and S' respectively, then the transformation formula (3-44) will be applicable regardless of whether the clocks in S and S' have been made to agree for some chosen event.

Let us now think of the displacement of a particle. Suppose that, in some given interval of time, the displacements relative to S and S' are s and s' respectively, and let the displacement of some fixed point in S' (say, the origin O' of a co-ordinate system chosen in it), relative to S be denoted by S , as in fig. 3-8. One then finds

$$s' = s - S, \quad (3-45)$$

which gives the transformation of displacement from one frame to another.

Problem 3-13

Establish formula (3-45) .

Answer to Problem 3-13

HINT: Make use of fig. 3-8. The proof is essentially similar to that of equation (3-41). Note that eq. (3-41) follows from equations (3-44) and (3-45).

One can, in this context, look at the transformation of the *separation* between two particles, as seen at any given instant of time. One can again make use of fig. 3-8, but now with a different interpretation where P and Q are to be interpreted as the positions of the particles at the same instant, and O'' is to be interpreted as being coincident with O' . One then gets

$$s' = s, \quad (3-46)$$

i.e., the vector representing the separation in S' is the *same* as that in S . Consequently, the magnitude of the separation, i.e., the scalar distance (or, in brief, the distance) also remains unchanged. Such quantities that remain unchanged in a transformation from one frame of reference to another are termed *invariants* of the transformation.

Equations (3-44) and (3-45) (or, alternatively, (3-41)) taken together are termed the *Galilean transformation* formulae.

Since the velocity of a particle is the derivative of the displacement with respect to time, the relation (3-41) is a consequence of (3-44) and (3-45).

3.9.4 Transformation of acceleration

The transformation formula for acceleration is obtained by making use of the definition (3-16b) together with the transformations of velocity and time (equations (3-41) and (3-44)):

$$a' = a - A. \quad (3-47)$$

Here a and a' are the instantaneous accelerations of a particle in the frames S and S' respectively, and A is the instantaneous acceleration of S' with respect to S . Once again it is assumed that S' does not possess a rotational motion relative to S . Notice that if the motion of S' is *uniform* relative to S , i.e., V is time-independent, then $A = 0$ and the accelerations in the two frames are equal:

$$a' = a. \quad (3-48)$$

We shall now make use of equations (3-47) and (3-48) to compare the equations of motion of a particle in different frames of reference.

3.10 Equations of motion in different frames of reference

3.10.1 Characterization of the inertial frames

Imagine a frame of reference S' to be *in uniform motion* relative to an *inertial* frame S . If a free particle is observed in S , then by definition of an inertial frame, its acceleration will be zero:

$$\mathbf{a} = 0. \quad (3-49a)$$

Then, by equation (3-48), the acceleration in S' will also be zero:

$$\mathbf{a}' = 0. \quad (3-49b)$$

In other words, the acceleration of a free particle in S' is zero, and so, by the definition of an inertial frame, S' is *also an inertial frame*.

This is an interesting result: *all frames uniformly moving relative to an inertial frame are inertial*. Indeed, this constitutes a complete characterization of the class of all inertial frames. A frame having a non-zero acceleration \mathbf{A} relative to an inertial frame cannot be inertial (check this out).

Put differently, *all inertial frames are in uniform motion relative to one another*.

3.10.2 Equations of motion of a particle in inertial frames

We have seen that the equation of motion of a particle in an inertial frame (say, S) is of the form

$$m\mathbf{a} = \mathbf{F}, \quad (3-50a)$$

where \mathbf{a} denotes the acceleration of the particle in S and \mathbf{F} is the force acting on it due to the influence of other particles or systems of particles. Considering all the individual

particles making up these systems with which the particle under consideration interacts, the force \mathbf{F} is determined, in the ultimate analysis, by the separations between the particle under consideration and all of these other particles, considered one at a time. But we have seen that the separations between any two particles, as seen in two different frames of reference, are the same (equation (3-46)). Consequently, the force acting on the particle is also the same, namely \mathbf{F} , in both S and S' , i.e.,

$$\mathbf{F}' = \mathbf{F}, \quad (3-50b)$$

where the primed symbols are used to denote physical quantities referred to S' .

Moreover, if S' be an inertial frame like S , then (3-48) implies that the acceleration \mathbf{a}' in S' is also the same as \mathbf{a} .

In non-relativistic mechanics, the *mass* of a particle is also independent of the frame of reference.

In the *relativistic theory*, however, the mass of a particle depends on its velocity and may thus be different as observed in various different frames of reference. This dependence of mass on velocity can be ignored when the velocity is small compared to the velocity of light in vacuum. The frames of reference in which the non-relativistic theory is applicable are such that in any of these, the particle velocity is indeed small in this sense, and it is in this sense only that the non-relativistic theory can be looked at as an approximation to the more complete and correct relativistic theory.

Thus, within the framework of the non-relativistic theory, one has

$$m' = m, \quad (3-50c)$$

where m and m' denote the mass of the particle in the two frames S and S' .

Collecting all this, one finds

$$m'\mathbf{a}' = \mathbf{F}', \quad (3-50d)$$

i.e., in other words, one has an equation of motion in S' *of the same form* as in S , where we recall that both S and S' have been assumed to be inertial frames of reference.

This is consistent with the fact that while stating Newton's second law we did not single out any particular inertial frame: the equation of motion (3-20) holds in any and every inertial frame.

Put differently, motions of mechanical systems, as described in different inertial frames, are *equivalent*. While the solutions to the equations of motion *look different* in the various different inertial frames, the equations of motion themselves are all of the same form. This is referred to as the *principle of equivalence* in Newtonian mechanics or, alternatively, the *Galilean principle of equivalence*.

3.10.3 Equation of motion in a non-inertial frame

We assume now that, while S is an inertial frame, S' possesses an acceleration \mathbf{A} relative to S and hence is a non-inertial one.

Moreover, S' will be assumed to possess only a translational motion with respect to S , and no rotational motion.

Here the equation of motion of a particle in S is of the form (3-20) where, in the present context, \mathbf{F} will be termed the *real* force on the particle, arising due to the influence on it exerted by other systems interacting with it.

Making use of equations (3-47), (3-50c) and the equality of \mathbf{F} and \mathbf{F}' , the latter representing the real force in S' , one then has

$$m'\mathbf{a}' = \mathbf{F}' - m'\mathbf{A}. \quad (3-51)$$

In other words, when S' is a non-inertial frame, the equation of motion in it is no longer of the form (3-50d), but involves an additional term $-m'\mathbf{A}$ in the right hand side. In a non-inertial frame, then, the rate of change of momentum does not equal the real force on the particle. Note that the additional term referred to above depends *only* on the frame S' (in addition to the mass of the particle), and not on any other mechanical system.

If, instead of S , one started from any *other* inertial system, the result (3-51) would have been the same since the acceleration of S' relative to that frame would also have been \mathbf{A} .

The term $-m'\mathbf{A}$ appearing in the equation of motion in a non-inertial frame is referred to as the *pseudo-force*, or *inertial force*, acting on the particle in that frame. Denoting this by the symbol \mathbf{G} , the equation of motion in a non-inertial frame is seen to be of the form

$$m\mathbf{a} = \mathbf{F} + \mathbf{G}, \quad (3-52)$$

where \mathbf{F} denotes the real force, and where we have dropped the prime appearing in the various symbols in (3-52). While we have arrived at (3-52) for a system possessing only translational motion relative to an inertial frame, in which case one has

$$\mathbf{G} = -m\mathbf{A}, \quad (3-53)$$

the form (3-52) remains valid for non-inertial frames that may possess rotational motion as well. In the latter case, however, the expression for the pseudo-force \mathbf{G} is of a different form involving, for instance, terms referred to as the *centrifugal force* and *Coriolis force* (see sec 3.21). These pseudo-forces do not make their appearance in the equation of motion in an inertial frame.

Problem 3-14

A small box of mass 1.0 kg rests on a smooth frictionless surface in a running train that is

decelerating at a rate of $0.5 \text{ m}\cdot\text{s}^{-1}$. Find the force that is to be exerted on the box to keep it at rest. Imagine the box to be replaced with a particle representing its center of mass (see sec. 3.16 and 3.17.3).

Answer to Problem 3-14

HINT: The acceleration of the train is $a = -0.5$ in its direction of motion (the negative sign is due to the fact that the train is decelerating); hence the pseudo-force on the box is $-ma = 1.0 \times 0.5$ (all quantities in SI units). Hence, in order to keep the box at rest, a force of 0.5n is to be applied on it in a direction *opposite* to the direction of motion of the train.

A few words are in order here regarding the quantitative definition of mass and force. In this, one has to turn to Newton's laws again. Indeed, while Newton's laws express a number of fundamental principles relating to the motion of systems of particles (we will come to Newton's third law in sec. 3.17.2 below), they provide, *at the same time*, operational procedures for the quantitative definition of mass and force. On the face of it, this might appear anomalous from the logical point of view. The important thing to realize, however, is that the laws are a set of formulations, derived from experience and observations, relating to the motion of particles and systems of particles. And the principal consideration relating to the validity of these laws is to see if they provide a number of well-defined predictions and operational procedures to check the correctness of these predictions in a broader sphere of experience and observations. This, more than mere logical merit, is what is necessary to make these laws useful in the physical sciences.

Making use of the laws in a number of situations involving the motion of particles, one can formulate the operational definition of mass and, in a similar manner, the operational definition of force can be obtained by employing these laws in some other specified situations. For this, one also needs a few other rules, once again formulated on the basis of experience like, for instance, the law of gravitation. What is then important to check is if, *all these taken together*, make up a theoretical structure that is internally *consistent* from the logical point of view and, at the same time, capable of explaining facts and phenomena observed in a *wider* sphere of experience. From this point of view, non-relativistic mechanics based on Newton's laws is, in an approximate

sense, an acceptable theory.

In this theory, mass and force are defined in terms of effective operational procedures. For the operational determination of mass one takes the help of an equipment known as the weighing balance, and follows a set of definite rules. And, for the determination of force, one measures the acceleration produced in a particle of known mass, or performs some related measurement. The question of logical soundness of this approach is one of some complexity. Without entering into this discourse we will, in this book, be content with the fact that the theory meets with necessary operational and practical requirements.

3.11 Force and work

In fig. 3-9, P and Q are any two points on the trajectory of a particle. Imagine the part of the trajectory between P and Q to be divided into a large number of small segments by the points R_1, R_2 , etc. Each segment being small in extent, it may be taken to be part of a straight line, and the force on the particle (\mathbf{F}) may be taken to be a constant throughout the segment. In the figure, one such segment has been shown in magnification, the end points of this segment being, say, A and B. Supposing that the position vector of A with reference to some chosen origin, say, O (not shown in the figure) is \mathbf{r} and the force acting on the particle at this position is $\mathbf{F}(\mathbf{r})$, one can assume that the force on the particle throughout the segment from A to B is $\mathbf{F}(\mathbf{r})$ (or, in brief, simply \mathbf{F} as shown in the figure; it is to be kept in mind, however, that the force varies, in general, with position on the trajectory). Consider the expression

$$W_{AB} = \mathbf{F} \cdot \mathbf{s} = |\mathbf{F}||\mathbf{s}| \cos \theta. \quad (3-54)$$

where \mathbf{s} stands for the displacement from A to B, θ is the angle between \mathbf{F} and \mathbf{s} as shown in the figure, and the suffix 'AB' refers to the fact that the above expression pertains to the segment AB on the trajectory. Since \mathbf{s} denotes a small displacement, the quantity W_{AB} is also a small quantity. It is termed the *work* done on the particle by the force in the displacement from A to B. Notice that $|\mathbf{F}| \cos \theta$ is the scalar component of \mathbf{F} along the displacement, and thus W_{AB} is simply the product of the magnitude of the displacement and the component of the force, the latter being a quantity carrying a sign ('+' or '-') depending on whether the vector component of \mathbf{F} along \mathbf{s} is in the same direction as or is directed opposite to \mathbf{s} .

The scalar and vector components of a vector, say, \mathbf{A} along another vector, say, \mathbf{B} are defined as, respectively, $\mathbf{A} \cdot \hat{\mathbf{b}}$ and $(\mathbf{A} \cdot \hat{\mathbf{b}})\hat{\mathbf{b}}$, where $\hat{\mathbf{b}}$ is the unit vector along \mathbf{B} (see sec. 2.7).

If now one sums up the work done on the particle by the force (sometimes this is referred to in brief as the work done on the particle, or the work done by the force, or even simply

as the work, provided the meaning is clear from the context) for all the small segments referred to above, then one obtains the work done on the particle in its motion from P to Q:

$$W = \sum \mathbf{F} \cdot \mathbf{s}. \quad (3-55)$$

In the above expression the summation is over a large number of tiny segments. The displacement in each segment is to be assumed infinitesimally (i.e., vanishingly) small. In order to indicate this one commonly uses, instead of \mathbf{s} , the symbol $\delta\mathbf{s}$ or, more commonly, $\delta\mathbf{r}$, where the latter denotes the vector directed from one end (\mathbf{r}) of the segment to the other ($\mathbf{r} + \delta\mathbf{r}$). The above summation then assumes the form $\sum \mathbf{F} \cdot \delta\mathbf{r}$ which, in mathematical terms, can be interpreted as an *integral*. It is referred to as the *line integral* of the force from P to Q (recall the definition of line integral in sec. 2.14.2) and is denoted as

$$W = \int_P^Q \mathbf{F} \cdot d\mathbf{r}. \quad (3-56)$$

1. I do not enter here into questions relating to the use of symbols like $\delta\mathbf{r}$ and $d\mathbf{r}$. While the former can be interpreted as a small, though finite, increment in \mathbf{r} , the latter is meaningful only under an integral sign. In this book, I shall often use the two symbols interchangeably.
2. While (3-56) gives the work done *by* the force-field under consideration, the quantity $-W$ is referred to as the work done *against* the force field.

As a special instance of the expression of work, consider the case where the displacement from P to Q takes place along a straight line, as in a one dimensional motion and where, moreover, the force \mathbf{F} on the particle is the same everywhere in this segment. One then has, from (3-56),

$$W = \mathbf{F} \cdot \mathbf{s}, \quad (3-57)$$

where \mathbf{s} stands for the displacement along the straight line from P to Q.

1. There may be, besides F , *other* forces acting on the particle at the same time. Indeed, in the above example, if F is not directed along s , then there has to be such a force in order that the particle can move along a straight line. The work done *by the force* F , however, is given by (3-56) or (3-57) independently of these other forces. Any other force acting on the particle will also have a work associated with it.
2. As I have already mentioned, the force is in general a function of the position (r), and the vector function $F(r)$ represents a *field of force*, an instance of a *vector field* (see sec. 2.13). In the special case when F is independent of the position r , one has a *uniform* field of force. In addition to the position, the force may also depend explicitly on *time*, in which case one talks of a *time-dependent* field of force, which may or may not be a uniform one. An example of a time-dependent force is found in the *forced oscillations* of a system.

The unit of force in the SI system being the newton (N), and the unit of displacement being the meter (m), the unit of work is N·m, which is commonly referred to as the *joule* (J).

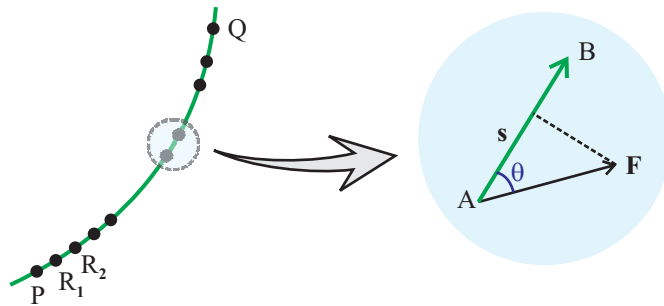


Figure 3-9: Illustrating the concept of work; the trajectory of a particle passes from P to Q through intermediate points R_1 , R_2 , etc., which divide it into a large number of tiny segments; one such segment (circled) is shown in magnification, where A and B are the end-points of the segment, which can be taken to be a straight line; s denotes the displacement from A to B, and F is the force at A (or, equivalently, at any other point on AB since the latter is vanishingly small), the angle between the two being θ .

Problem 3-15

(A) Consider the motion of a particle in a plane on which there acts a constant force $\mathbf{F} = \hat{i} + \hat{j}$, where \hat{i} and \hat{j} are unit vectors along the x- and y-axes of a Cartesian co-ordinate system in the plane. Suppose the particle moves along a straight line from the origin O to a point P with co-ordinates (2, 0), and then similarly from P to Q with co-ordinates (0, 2). Find the work done by the force in these two parts of the journey. Compare the sum of the two results with the work done by the same force in a rectilinear motion of the particle from O to Q. All physical quantities are assumed to be in SI units. (B) Consider now a position dependent force $\mathbf{F}(\mathbf{r}) = kr^3\hat{r}$ where k is a constant and \hat{r} denotes the unit vector directed along the position vector of a point in the plane. Obtain the work done in each of the three motions mentioned in (A), and make the same comparison. .

Answer to Problem 3-15

HINT: (A) All the three motions here are along straight lines and the force acting on the particle is a constant one. Hence, in accordance with eq. (3-57) the work (in J) from O to P is $W_1 = (\hat{i} + \hat{j}) \cdot (2\hat{i}) = 2$, that from P to Q is $W_2 = (\hat{i} + \hat{j}) \cdot (-2\hat{i} + 2\hat{j}) = 0$, while that from O to Q is $W_3 = (\hat{i} + \hat{j}) \cdot (2\hat{j}) = 2$. Thus, $W_1 + W_2 = W_3$.

(B) Note that the unit vector \hat{r} for any given point (x, y) in the plane makes an angle θ with the x-axis, where $\tan\theta = \frac{y}{x}$, and hence $\hat{r} = \frac{x}{\sqrt{x^2+y^2}}\hat{i} + \frac{y}{\sqrt{x^2+y^2}}\hat{j}$. As the particle moves a short distance from, say, (x, y) to $(x + dx, y + dy)$, the work done by the force is

$$\mathbf{F} \cdot d\mathbf{r} = F_x dx + F_y dy = kx(x^2 + y^2)dx + ky(x^2 + y^2)dy.$$

In the motion from O to P, one has $y = 0$, $dy = 0$, and thus $W_1 = \int_0^2 kx^3 dx = 4k$. Similarly, in the motion from P to Q, $y = 2 - x$, $dy = -dx$, and thus $W_2 = k \int_2^0 (2x - 2)(2x^2 - 4x + 4)dx = 0$, and finally, $W_3 = 4k$. Thus, once again, $W_1 + W_2 = W_3$. Indeed, both the force fields in (A) and (B) are conservative (see sec. 3.15.2).

Power.

The rate at which a force (or an agent exerting a force) performs work, is termed *power*. Thus, if 10 J of work is performed at a uniform rate in 2 s, the power exerted is $5 \text{ J}\cdot\text{s}^{-1}$ or, in brief, 5 *watt* (5 W), where the watt ($\text{W} \equiv \text{J}\cdot\text{s}^{-1}$) is the SI unit of power.

3.12 Work in rectilinear motion. Potential energy.

If the motion of a particle takes place along the x -axis of a Cartesian co-ordinate system and if, at any position of the particle corresponding to a co-ordinate, say, x , the force acting on it is F (evidently the resultant force acting on the particle has to be along the x -axis; for the time being there is no harm in taking F as this resultant force; alternatively, for a force pointing in any other direction, F can be taken as the x -component of the force) then the work done on the particle in a displacement δx is given by the expression $F\delta x$.

In this case, considering a reference point with co-ordinate, say X , the expression for work done in a displacement from X to x is obtained from (3-56) as the integral

$$W = \int_X^x F(x)dx. \quad (3-58)$$

The symbol x inside the integral need not be confused with the same symbol used as the upper limit of integration. The former is a dummy integration variable and can be replaced by any other appropriate symbol, say x' .

Assuming that the force F depends only on the position of the particle and not on time, the above integral depends only on the final co-ordinate x of the particle, provided the reference point X is fixed. This means that the work done by the force on the particle in a displacement from the fixed reference position X to any given position x is uniquely determined by x alone, i.e., in other words, is some definite function of x . This function, taken with a *negative* sign, i.e., the work done *against* the force in the displacement from X to x , is termed the *potential energy* of the particle at the position x (with reference, one may add, to the point X). Denoting this by the symbol $V(x)$, one obtains the following mathematical definition of potential energy

$$V(x) = - \int_X^x Fdx. \quad (3-59)$$

The potential energy $V(x)$ defined above is of considerable relevance in the motion of a

particle, where we are considering only one dimensional motions in the present section. In particular, the following observations are of relevance.

1. For a given reference point X , the potential energy depends only on the position x of the particle, and not on the path of integration from X to x . For instance, in fig. 3-10, the potential energy calculated for a path from A (the reference point) to P (the point with co-ordinate x) along the shortest distance will be the *same* as the one calculated for a path from A to B and then from B to P.
2. Equation (3-59) tells us that the potential energy is the integral of the force, taken with a negative sign. Conversely, then, the force can be obtained as the *derivative* of the potential, once again with a negative sign:

$$F(x) = -\frac{dV}{dx}. \quad (3-60)$$

3. The equation of motion of a particle, and hence the nature of its motion, is determined by the field of force $F(x)$ acting on it. The question that comes up is, does the field of force determine the potential energy (*potential*, in short) uniquely? The expression (3-60) tells us that, adding a constant term to $V(x)$, leads to the *same* expression for $F(x)$. In other words, two different potential energy functions $V(x)$ and $V'(x) = V(x) + V_0$, where V_0 is a constant, correspond to the same field of force $F(x)$ and one can use either expression for the potential energy of the particle. This non-uniqueness is actually related to the arbitrariness in the choice of the reference point X . Indeed, choosing the reference point as, say, X' instead of X , one obtains, from (3-59) a potential energy $V'(x)$ that differs from $V(x)$ by a constant, independent of x .
4. When talking of the work done on the particle from X to x (here, as elsewhere, we use an abbreviation like, say, 'the point x ', or even, simply, ' x ', in place of the longer expression 'the point with co-ordinate x '), it is implied that an actual motion from X to x along some path takes place. On the other hand, in the expression (3-59), the point X is simply a reference point used as the lower limit in the integration, because $V(x)$ primarily refers to the point x , and the expression for $V(x)$ does not

necessarily imply an actual motion from X to x .

5. The difference in potential energy for any two points x_1 and x_2 is given by

$$V(x_2) - V(x_1) = - \int_X^{x_2} F dx + \int_X^{x_1} F dx = - \int_{x_1}^{x_2} F dx. \quad (3-61)$$

Notice that this difference does not depend on the choice of the reference point X . Choosing the reference point at a point X' instead of X alters the value of potential energy, but the change, being a constant independent of x , *cancels out* in the difference of potential energy between any two points.

6. Equation (3-61) tells us that the change in potential energy of a particle in rectilinear motion from x_1 to x_2 is the work done on it, taken with a *negative* sign (check against eq. (3-58)).
7. As we will see, a potential energy can be defined for motions in two- and three dimensions as well, *provided the force-field satisfies certain conditions*. A force field for which a potential energy function can be defined is referred to as a *conservative* one. In contrast to two- and three dimensional motions, a force-field in one dimension is always conservative.



Figure 3-10: A reference point (A) and two other positions (P and B) for a particle in one dimensional motion; the potential energy worked out for a path from A to P along the shortest distance is the same as that worked out for a path from A to B and then from B to P.

Gravitational potential energy.

As an example of potential energy in one dimensional motion, consider a particle under the action of earth's gravitational pull where, for the sake of simplicity, we assume that the motion is confined to a vertical line. The force on the particle is mg , where m is the mass of the particle and g is the *acceleration due to gravity* (see sec. 5.4.1) which, to a good degree of approximation, can be taken to be a constant. The force of gravity acts

in the vertically downward direction, which we take to be the negative direction of the z -axis of a Cartesian co-ordinate system.

Assuming the origin ($z = 0$) to be on the surface of the earth, which we take as the reference point for the calculation of potential energy, it is now a simple matter to work out the potential energy of the particle at any given height, say, h from the earth's surface. Making use of eq. (3-59), one gets

$$V(h) = - \int_0^h (-mg) dz = mgh. \quad (3-62)$$

This represents the work to be performed *against* gravity in raising a body of mass m through a height h above the earth's surface.

While we have considered here the restrictive case of a rectilinear motion, the expression for the potential energy remains valid even when one considers the more general case of three dimensional motion under the earth's gravitational attraction. The field of force resulting from the gravitational pull of the earth on a particle is a *conservative* one (see sec. 3.15.2), the force on a particle of mass m at any given point being $-mg\hat{k}$ where \hat{k} stands for the unit vector in the vertically upward direction, and where it is assumed that the point under consideration is sufficiently close to the surface of the earth and its motion is restricted to a region whose linear dimension is small compared to the earth's radius. The potential energy of the particle then depends only on its z -co-ordinate, and is given by the expression (3-62) for a point with co-ordinate h .

3.13 Kinetic energy

The kinetic energy of a particle is defined as (refer to sec. 3.2.2.7)

$$K = \frac{1}{2}mv^2, \quad (3-63)$$

where m is the mass of the particle, while its velocity is \mathbf{v} (i.e., $v^2 = \mathbf{v} \cdot \mathbf{v}$).

This definition has a considerable significance in mechanics, which we shall now have a

look at. I shall then introduce the principle of conservation of energy - a principle of vast importance in physics. In the present section, the principle of conservation of energy will be established for the rectilinear motion of a particle under a constant force, while the question of validity of the principle under the more general condition of motion in three dimensions under a *conservative* field of force will be taken up in sec. 3.15

In the rectilinear motion of a particle, all the quantities like displacement, velocity and force on the particle are directed along a fixed straight line which one can take as the x-axis of a Cartesian co-ordinate system. If P and Q be two positions of the particle on this straight line, for which the position co-ordinates and velocities are respectively x_P , x_Q , and v_P , v_Q , then the displacement from P to Q is

$$s = x_Q - x_P, \quad (3-64a)$$

while the kinetic energies at the two positions are

$$K_P = \frac{1}{2}mv_P^2, \quad K_Q = \frac{1}{2}mv_Q^2. \quad (3-64b)$$

Suppose first that the force on the particle is constant, independent of its position. This means that the acceleration a is also independent of x . Then, taking P as the initial position and Q as the final one, one has, from eq. (3-35)

$$v_Q^2 = v_P^2 + 2as, \quad (3-65a)$$

i.e., in this case,

$$K_Q = K_P + mas, \quad (3-65b)$$

Making use of the equation of motion of the particle, one can write $ma = F$, the force on the particle. Since, in the present instance, the force is a uniform one and acts in the same direction in which the displacement takes place, the product Fs is nothing but the

work done on the particle in its displacement from P to Q. Denoting this by W , one has

$$W = K_Q - K_P, \quad (3-66)$$

which tells us that *the work done on the particle is equal to the increase in its kinetic energy*. This, then, is where the significance of what we have defined as the kinetic energy lies: *the increase in kinetic energy equals the work done on the particle*. As we will see later, this statement holds in more general contexts as well.

3.14 Principle of conservation of energy in rectilinear motion

We have already seen that the increase in the *potential* energy of the particle equals the work done *against* the force:

$$-W = V_Q - V_P. \quad (3-67)$$

Equations (3-66) and (3-67) taken together imply

$$K_Q + V_Q = K_P + V_P. \quad (3-68)$$

In other words, considering any two points in the path of motion of the particle, the sum of kinetic and potential energies will be the same for both the points. Calling this sum the *total energy* of the particle, one can thus say that this total energy remains a constant during the motion of the particle. This is the *principle of conservation of energy* in the present context, i.e., for the motion of a particle in one dimension under a constant force.

Problem 3-16

Consider a particle moving along a straight line under the action of a force for which the potential

energy is $V(x)$. Set up its equation of motion and identify its *equilibrium* position(s) where an equilibrium position is one where the particle, once at rest, continues to stay at rest.

Answer to Problem 3-16

SOLUTION: Making use of eq. (3-60) in the equation of motion

$$ma = m \frac{d^2x}{dt^2} = F(x), \quad (3-69a)$$

one obtains, in terms of the potential energy,

$$ma = m \frac{d^2x}{dt^2} = -\frac{dV}{dx}. \quad (3-69b)$$

Consider a point x_0 at which the acceleration of the particle is zero. This means that if the particle be located at rest at x_0 at any given instant of time then it will continue to be at rest at all times (reason this out; consider a succession of infinitesimally small time intervals), i.e., in other words, x_0 is an equilibrium position for the particle. Put differently, an equilibrium position is given by

$$\frac{dV}{dx} = 0. \quad (3-70)$$

For motion under a given potential, there may be more than one equilibrium positions (or, possibly, none, as in the case of motion under a constant force) for the particle.

3.15 Kinetic energy, potential energy, and work

3.15.1 Kinetic energy and work

In the last section we have looked at the relation between the kinetic energy of a particle and the work done on it, for a one dimensional motion under a constant force. In the present section we will consider the more general situation of three dimensional motion of a particle under a field of force that is not necessarily a uniform one.

The concept of work done by a field of force on a particle in three dimensional motion has already been explained (see eq. (3-55) and the paragraph leading to eq. (3-56)). The definition of the kinetic energy of the particle, as given in eq. (3-63), is also valid for three

dimensional motion. The interesting connection to note in the context of these two is that, the relation between the work done on the particle and the increase in its kinetic energy, expressed in eq. (3-66), *remains valid for three dimensional motion as well*. In other words, if P and Q denote two positions of a particle in its three dimensional motion along a trajectory under a field of force that is not necessarily a uniform one, one will still have

$$W = K_Q - K_P, \quad (3-71)$$

where W is the work done on the particle in its motion from P to Q.

Problem 3-17

Derive eq. (3-71), taking the cue from the way the corresponding relation was derived for one dimensional motion under a uniform force (eq. (3-66)) .

Answer to Problem 3-17

HINT:

$$W = \int_P^Q m \frac{d\mathbf{v}}{dt} \cdot d\mathbf{r} = \int_P^Q m \frac{d\mathbf{r}}{dt} \cdot d\mathbf{v} = \int_P^Q m \mathbf{v} \cdot d\mathbf{v} = \frac{m}{2} (v_Q^2 - v_P^2).$$

3.15.2 Potential energy. Conservative force-fields.

We now address the question of defining the potential energy in three dimensional motion. The important observation to make here is that, unlike the concept of kinetic energy, the *potential energy function cannot, in general, be defined for three-dimensional motion*. Indeed, the possibility of defining a potential energy function depends on whether the force field is a *conservative* or a *non-conservative* one.

Imagine any two points P and Q in a field of force and any path or line connecting these two points, which *need not be* a trajectory of the particle connecting the two points P and

Q. Among the innumerable possible paths connecting P and Q, two are shown in fig. 3-11. Considering any of these possible paths, one can imagine it to be divided into a large number of short segments, in a manner similar to that indicated in 3-9. Referring then to a typical segment as shown in magnification in fig. 3-9, one can consider the quantity $\mathbf{F} \cdot \delta \mathbf{r}$ where \mathbf{F} stands for the force on the particle at any point in the segment (it does not matter which point, since the segment is, in the end, assumed to be vanishingly small), and $\delta \mathbf{r}$ stands for the vector length of the segment, i.e., the vector extending from the initial to the final point (denoted by s in fig. 3-9, showing a segment in magnification). If one now adds up all these quantities for the various segments of the path, and goes over to the limit where the segments are of vanishingly small length one obtains the *line integral* of the force along the chosen path from P to Q:

$$\sum \mathbf{F} \cdot \delta \mathbf{r} \rightarrow \int_{(\text{path})} \mathbf{F} \cdot d\mathbf{r}. \quad (3-72)$$

This sounds exactly like a repetition of how we arrived at (3-56). While there is indeed much in common in the way the expressions (3-72) and (3-56) have been arrived at, the two expressions differ in significance. The latter stands for the *work done* on the particle under consideration in a possible motion of it from an initial to a final point. The former, on the other hand, is the *line integral* of the force field along *any* conceivable path, not necessarily the path followed in a possible motion of the particle. For a path describing a possible motion (for instance, the path marked 1 in fig. 3-11), the two reduce to the same thing.

Now comes the big condition. Suppose that the above line integral of the force field is *path independent*, i.e., the value of the integral is the same whatever the path chosen - same, for instance, for the two paths shown in fig. 3-11. The line integral is then determined solely by the two points P and Q chosen in the force field. If this condition of path independence applies for *all* pairs of points like P and Q, one says that the force field is *conservative* in nature. Else, it is said to be a *non-conservative* one.

Given the function $\mathbf{F}(\mathbf{r})$ defining the force-field under consideration, there are ways to

tell whether or not the field is a conservative one according to the above definition, but I will not go into that.

The important result in the present context is that, *if* the force field under consideration happens to be conservative, then one can define a potential energy function V such that the above line integral can be expressed as the difference in the potential energy values at P and Q:

$$\int_P^Q \mathbf{F} \cdot d\mathbf{r} = V_P - V_Q. \quad (3-73)$$

Here I have introduced a slight difference in the notation for the line integral compared to eq. (3-72), emphasizing that its value depends only on the end points P and Q. This equation itself provides a definition of the potential energy if one of the points P and Q is chosen as a reference point. For instance, choosing Q as the reference point, the potential energy at the point P, with position vector, say, \mathbf{r} , is obtained as

$$V(\mathbf{r}) = - \int_{\mathbf{r}_0}^{\mathbf{r}} \mathbf{F} \cdot d\mathbf{r}, \quad (3-74)$$

where \mathbf{r}_0 is the position vector of the reference point Q, and the integration is performed over any path connecting P and Q. This definition involves an arbitrariness of an additive constant in the value of $V(\mathbf{r})$, namely $V(\mathbf{r}_0)$, which we have taken to be zero here.

In practice the force fields acting on a particle or a system of particles are such that the force falls off to zero at infinitely large distances. For such a force field, the reference point is commonly taken to be a point located at an infinitely large distance, and is referred to as ‘the point at infinity’. The idea underlying such a choice will be discussed at greater length in the context of gravitation and electrostatics in sections 5.2.4 and 11.4.4. Incidentally, the term ‘potential’ is commonly used in mechanics as a shortened version of ‘potential energy’. In gravitation and in electrostatics, however, the term ‘potential’ is used in a slightly different sense, namely the potential energy per unit mass and per unit charge respectively.

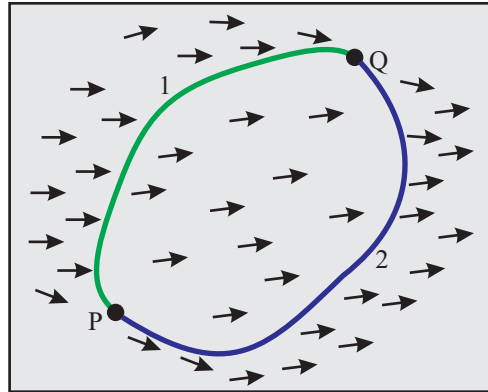


Figure 3-11: Two paths, marked 1 and 2, connecting points P and Q in a force-field, depicted with the help of arrows drawn at various points; the path marked 1 represents a trajectory of the particle under the force-field; for a conservative force-field, the line integral of the force is the same for the two paths, being equal to the work done by the force in a motion from P to Q.

3.15.3 Principle of conservation of energy: the general context

Now consider a possible trajectory connecting P and Q, i.e., a path followed by the particle in an actual motion from P to Q under the given force field (path marked 1 in fig. 3-11). In accordance with eq. (3-56), the value of the integral gives, by definition, the work done on the particle by the force field in its motion from P to Q. Denoting this by W , one obtains

$$V_Q - V_P = -W. \quad (3-75)$$

i.e., in other words, the change in potential energy from P to Q equals the work done *against* the force-field in a possible motion of the particle from P to Q (for the other path, marked '2' in the figure, the value of the path integral remains the same, but it cannot be interpreted as work done if the path does not represent a possible motion). Combining with eq. (3-71), one arrives at the *principle of conservation of energy*: the sum of kinetic and potential energies is constant along any given trajectory in a conservative force-field.

Put differently, the *total* energy of a particle remains constant during its motion in a conservative force-field, remaining unchanged at its initial value, i.e., the value at any

chosen point of time, say $t = t_0$.

3.15.4 Work and energy: summary

The work done by a field of force on a particle moving in three dimensions is given by the expression (3-56) where P and Q denote the initial and final positions of the particle on its trajectory, and where the terms ‘initial’ and ‘final’ refer to any two time instants of choice. A one dimensional motion is a special case of the more general three dimensional motion, and the expression of work done reduces to (3-58) where X and x denote the co-ordinates of the initial and final positions of the particle. If, moreover, the force on the particle is a uniform one, independent of the position and the displacement takes place along a straight line, the expression for work done assumes the simpler form (3-57).

The kinetic energy of a particle is defined by eq. (3-63). Its significance lies in the fact that the increase (here the term increase is used in the algebraic sense, i.e., with the appropriate sign) in kinetic energy between any two points P and Q on a trajectory of the particle is equal to the work done by the field of force from the initial to the final point along the trajectory.

The potential energy function $V(x)$ for a rectilinear motion is defined by the expression (3-59), which implies that the increase in potential energy is equal to the work done by the force, taken with a *negative* sign (eq (3-61)). Taken together with the expression for increase in kinetic energy, this implies the principle of conservation of energy in a rectilinear motion.

For a particle in three dimensional motion, *one cannot define a potential energy unless the force field is conservative*. The condition for the force field to be conservative is that the line integral of the force between any two given points is to be path independent, depending only on the two points chosen. For such a field of force, the principle of conservation of energy holds once again, where the potential energy is defined by eq. (3-74). While it may appear that the condition for a force field to be conservative is a restrictive one, conservative force fields are still of great relevance and importance in physics.

A one dimensional motion, where both the force and displacement are along the same direction, say, the x-axis, is simply a special case of three dimensional motion where the condition of the line integral of the force to be a function of the end points alone is automatically met with, provided one considers paths of integration restricted to the x-axis alone. In this case the definition of potential energy and the validity of the principle of conservation of energy does not require any additional condition on the field of force.

It may be noted that, while the energy is constant along any given trajectory, its value differs, in general, from one trajectory to another. The expression for energy is given by

$$E = \frac{1}{2}mv^2 + V(\mathbf{r}), \quad (3-76)$$

which is a function of the instantaneous position and velocity of the particle. Particles moving under various different conservative force fields are characterized by different potential energy functions $V(\mathbf{r})$. For a given force-field, there may exist, apart from the energy defined in eq. (3-76), *other* functions of position and velocity, that remain constant along any given trajectory of the particle. Such functions are termed *constants of motion* in mechanics.

Finally, let the forces $\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_n$ act simultaneously on a particle in a displacement from, say P to Q. Then the work W performed on the particle is equal to the work that would be performed by a single force $\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2 + \dots + \mathbf{F}_n$ (i.e., by the resultant of the given forces) in the same displacement from P to Q (reason this out: the expression for work involves the force linearly).

1. Phrases like 'work done by a force on a particle' or 'work done against a force' may at times cause confusion. Considering, for instance, two systems, say, A and B, the constituent particles of B exert forces on those of A while similarly, particles in A exert forces on those in B. The total work done by forces exerted by particles in B on those of A in a displacement of the latter will be termed the work done by B on A. The work done by the system A on the system B is similarly defined. Thus, in the ultimate analysis, work is performed by a particle or a system of particles on another and, in this sense, force is a means of performing work. Keeping this

in mind, one can avoid confusion as to who performs work on what.

2. The concept of work being done by a particle or a system of particles on another needs to be extended by treating the *electromagnetic field* as a *dynamical system*, on the same footing as a system of particles. This I briefly discuss in section 3.15.6 in connection with a broader interpretation of the principle of conservation of energy.
3. Finally, consider a situation in which a particle, say, A is acted on by forces F_1 and F_2 due to two other fixed bodies, say B and C. If the two forces act simultaneously, then the work done on A in any given motion is given by a sum of two expressions, each of the form (3-56) (with F_1 and F_2 respectively replacing F), where the integrals are taken along the path followed by the particle under the joint action of the forces.

If the forces F_1 and F_2 are conservative in nature then something more can be said. Suppose that Γ_1 is a path followed by the particle in moving from the point P to the point Q under the action of F_1 alone, and similarly, Γ_2 is a path followed between the same two points under the action of F_2 alone. Let the amounts of work done in the two motions be W_1 and W_2 respectively. Assume now that, under the simultaneous action of the two forces, the particle A follows a path Γ in moving between the same two points. Then the work done in this motion (this work can be said to be done by the *resultant* of the two forces F_1 and F_2) will be given by $W_1 + W_2$. This is an expression of the *independence* of the forces acting simultaneously on a particle, exerted by various different bodies. In the ultimate analysis, it is a consequence of the fact that known forces in nature can be described as *two-body* forces.

Problem 3-18

A particle of mass 0.02 kg moves under the action of a conservative force field in a region in which P is a reference point at which the potential energy is assumed to be zero, while Q and R are two other points. In a motion of the particle starting from rest at P and reaching R through Q, its speed is $1.0 \text{ m}\cdot\text{s}^{-1}$ at Q and $1.5 \text{ m}\cdot\text{s}^{-1}$ at R. In another motion, the particle starts from Q with a speed of $2.5 \text{ m}\cdot\text{s}^{-1}$ and reaches R. What will be its speed at R ? .

Answer to Problem 3-18

Since, in the first motion, the total energy at P is zero and this energy is conserved, the potential energies at Q and R are, respectively, $-\frac{1}{2} \times 0.02 \times (1.0)^2$ J and $-\frac{1}{2} \times 0.02 \times (1.5)^2$ J. Thus, the total energy in the second motion is $E = -\frac{1}{2} \times 0.02 \times (1.0)^2 + \frac{1}{2} \times 0.02 \times (2.5)^2$ J. If the required speed at R be v m·s⁻¹ then, invoking the conservation of energy for the second motion, one has $\frac{1}{2} \times 0.02v^2 = E - (-\frac{1}{2} \times 0.02 \times (1.5)^2)$. This gives $v = 2.74$ (approx).

3.15.5 Energy as ability to perform work

When a system performs work on another, the dynamical property of the system that makes the performance of work possible is its *energy*. In a sense, the energy of a system is a measure of its ability to perform work on other systems. Let us see what this means by referring to the potential energy and kinetic energy of a single particle.

Suppose that a particle moves under a given field of force from a reference point P to the position Q, along the trajectory represented by the path marked 1 in fig. 3-11. If, in this displacement, the work done on the particle by the force-field be W then the potential energy of the particle in position Q relative to that at P is given by (see eq. (3-75))

$$V = -W. \quad (3-77)$$

Suppose now that the particle is taken back from Q to the reference position P, in which process R (not shown in fig. 3-11) represents a typical intermediate position, the force on the particle in this position due to the force-field under consideration being, say, F , where the latter generates an acceleration in the particle. Imagine, however, that an equal and opposite force is being exerted by some other system (call it S) at all such intermediate positions so as to cancel the acceleration of the particle. As a result, the velocity and hence the kinetic energy of the particle remains unchanged at its value at the position Q while, at the same time (a) the potential energy of the particle in the field of force under consideration is reduced by V , and (b) the particle performs work on S by the *same* amount V .

This requires that the interaction or coupling between the particle and the system S be set up in a specific manner, namely one in which there is no change in the *mutual* potential energy between them. An interaction between the particle under consideration and the system S, appropriately designed, is possible in principle where the above consequences ((a) and (b)) are realized.

In other words, a particle can be said to have an ability to perform work on a system by virtue of its position (Q in the present instance) in a given field of force, and the amount of work that it can perform equals its potential energy in that position due to the field of force. In this context, one needs to specify a reference point (P in the present instance) in order that the potential energy and the work may have specific values.

We now look at a similar interpretation of the kinetic energy of the particle. Imagine the particle to be located at Q with kinetic energy K, where one need not now consider the field of force that was originally responsible in imparting this kinetic energy to the particle. Suppose that the particle is made to interact with a system S in such a way that the particle eventually comes to rest due to the force exerted on it by S. The work done by *this* force on the particle is then (compare eq. (3-71) with an appropriate reinterpretation of symbols)

$$W' = -K. \quad (3-78)$$

Assuming that the interaction or coupling between the particle and the system S is of an appropriate kind (wherein there is no change in the mutual potential energy between them), the work done by the particle on the system S is found to be

$$W = -W' = K. \quad (3-79)$$

One thus finds that the kinetic energy of the particle with any given velocity is the same as the work it performs on a system (S), appropriately coupled to it, before it is made to come to rest by the force exerted on it by the latter.

Collecting the above results one can then make the following statement. Consider a particle in a conservative field of force, at the position Q with a velocity v , for which its total energy is, say, $E = V_Q + K_Q$. Now assume that the particle is made to interact with a given system S , where the interaction is of a certain specific type such that there is no change in the mutual potential energy between the particle under consideration, and the system S . Let, under this interaction, the particle be made to traverse a closed trajectory whereby it ends up at Q , but now with velocity zero. Let the work done by the particle on S in this process be W_1 . Assume next that The particle is made to interact with a second system S' , where the interaction is again of the above type, but is such that the particle is brought from Q to an appropriate reference position P (the same position with respect to which is potential energy is defined), while maintaining its velocity at zero value. Let the work done by the particle on S' in this process be W_2 . The *total* work ($W_1 + W_2$) done by the particle in the combined process (which brings it to the reference position and reduces its velocity to zero) is then the same as its energy E in the initial state (position Q , velocity v).

This result, which we have arrived at for a single particle, can be extended to a system of particles as well.

3.15.6 Conservation of energy: a broader view

Note that what we mean by the conservation of energy is the constancy of the sum of kinetic and potential energies of a particle. However, the term ‘energy’ has a wider connotation in physics and the principle of conservation of energy is more far-reaching in scope. Starting from the equation of motion of a single particle we shall, in subsequent sections, look at the dynamics of a system of particles, where we will see that the concepts of energy and conservation of energy continue to be relevant. Further, the two concepts retain their relevance in the case of thermodynamic systems made up of *large* numbers of constituents, though one needs for such systems an appropriate accounting procedure for energy.

As an extension of the concept of energy, one may have a look at the principle of *equiva-*

lence of mass and energy. While our principal concern in this book will be a presentation of the basics of *non-relativistic* mechanics (relativistic principles will be briefly outlined in chapter 18), it is the *relativistic* theory that has a wider applicability. Indeed, the non-relativistic theory can be looked upon as deriving from the more solidly founded relativistic theory through a certain approximation scheme. In the relativistic theory, the mass of a particle is dependent on its velocity.

In this context one has to distinguish between the *rest mass* and the *moving mass* of the particle: the mass as observed in a frame of reference in which the particle is at rest is termed the rest mass, while the mass observed in any other moving frame of reference is referred to as the moving mass. It is the latter that depends on the frame of reference, i.e., on the velocity of the particle. For a system of particles, the sum of the rest masses of the constituent particles is taken as the rest mass of the system, though there may not be any frame of reference where all the particles are simultaneously at rest.

Though the rest mass of a particle or a system of particles is an intrinsic characteristic, it is however, not absolute or immutable. Under certain circumstances, the system under consideration may perform work on another system *at the expense of its rest mass*. One can then say that a certain amount of *rest mass has been converted into energy*. This statement becomes more meaningful by the observation, corroborated from experimental evidence, that there exists a relation of proportionality between the change in rest mass and that in energy. Indeed, one finds that a consistent and meaningful definition of the energy of a particle requires it to be related to its *moving* mass (m) as

$$E = mc^2, \tag{3-80}$$

where c stands for the velocity of light in vacuum. One can then say, from a broader point of view, that m and E signify one and the same physical quantity - the fact that their conventional definitions do not take into account this relatedness, has resulted in these being measured in different units and in the appearance of the factor c^2 in their relation. This identity of the two physical quantities is, from the fundamental point of view, the basis of the principle of equivalence of mass and energy.

This means that, when the conversion of mass into energy is taken into account, which includes, in particular, the conversion of rest mass into energy, the principle of conservation of energy acquires a broader scope and significance: a phenomenon where energy appears to be created and destroyed and the principle of conservation of energy appears to be violated, is in reality, found to *corroborate* the principle when the equivalence of mass and energy is taken into account.

This, on the face of it, leads to a paradox. On the one hand, the principle of conservation of energy requires the force-field to be a conservative one, implying that the principle holds only under rather restrictive conditions. At the same time, one finds that the scope of the principle of conservation of energy is a broad one, involving the possibility of conversion of mass into energy. How can this be reconciled with the restrictiveness imposed by the requirement of a conservative force field?

In reality, the concept of a conservative force field is, in a sense, more fundamental than that of a non-conservative one. If a system of particles interacts with other systems and if the effect of those other systems on it is considered only in an average sense, without regard to the details of the states of motion of those other systems, then this effect, in practice, can be described as a non-conservative force on the system under consideration. The approach of ignoring the detailed motions of a large number of particles and looking at the effect of these particles on a given system only in an average sense, is one imposed by practical necessity, and does not qualify as a fundamental description of the underlying dynamical principles. A more fundamental approach would be to include the motions of *all* the interacting particles in the theory while assuming at the same time that all the forces are conservative.

Assuming that the fundamental forces are all conservative in nature, energy conservation is implied by another basic principle, namely the principle of *homogeneity of time*, which tells us that the choice of the origin of time should not be of any consequence in formulating the basic laws of motion for a system. This provides the basis for the principle of conservation of energy for any *closed* system. For a system in interaction with a large number of other particles, even a very weak interaction would require that

the energy exchange with these external particles be taken into account. An instance of such energy exchange is provided by the flow of *heat* into or out of a system that has to be taken into account in establishing the principle of conservation of energy in thermodynamics in the form of the *first law of thermodynamics*.

In this book, the only non-conservative force in mechanics I will consider will be the force of *friction*. Even there, when the energy appearing as heat (or, more precisely, as internal energy) in the systems under consideration is taken into account, one finds that the principle of conservation of energy holds good.

Finally, I have so long been talking only of material bodies in the accounting of energy, assuming that energy is exchanged between material bodies. The material body may, for instance, be an extended, continuously distributed medium in which *elastic deformations* may take place. An energy is, in general, required for such a deformation. Such deformation energy is associated with the propagation of sound waves in a medium. From a fundamental point of view, however, the energy carried by means of sound waves can be interpreted as kinetic and potential energies of the constituent parts of the system, which can be looked upon as material bodies themselves.

However, the concept of energy exchange between material bodies also needs extension, and a more appropriate concept would be that of energy exchange between *dynamical systems*, where one includes the electromagnetic field (as also fields of other types like, possibly, the gravitational field) in the category of a dynamical system. For instance, when we observe the light emitted from a source, what actually happens is that an amount of energy coming out of the source is carried by the electromagnetic field in space and produces the sensation of vision on reaching our eyes. I shall include a brief introduction to electromagnetic fields and the energy associated with an electromagnetic field in chapter 14. A fundamental point of view of great importance in physics is that what we perceive as material bodies are, in the ultimate analysis, nothing but states of *fields* of various description. Energy conservation in this point of view then means the conservation of energies of these fields in interaction. But this is an area I cannot go into in this book (see sec. 16.15 for a brief introduction).

In these last few paragraphs I have digressed from the mechanics of a single particle to a consideration of the principle of conservation of energy for systems of more general description. We will now take a summary overview of a number of aspects of the mechanics of a particle and then move on to basic principles underlying the mechanics of a *system of particles* and, as an important special case, the mechanics of a *rigid body*.

Analogous to the transformations of physical quantities such as velocity and acceleration of a particle in a transformation of the frame of reference, one can think of the transformations of kinetic energy and potential energy as well. The transformation of kinetic energy can be worked out from that of the velocity, while the transformation of the potential energy depends on the inertial forces in the frames of reference concerned. The principle of conservation of energy remains valid in either of the following two forms: (a) the work done on a particle in its motion from an initial to a final instant of time equals the change in its kinetic energy; and (b) the sum of kinetic and potential energies for a conservative system remains constant during the motion.

3.16 Mechanics of a single particle: overview

The motion of a single particle is explained in terms of its equation of motion which depends on the frame of reference chosen. In an inertial frame the equation of motion relates the rate of change of momentum to the ‘real’ force acting on it (equation (3-20), (3-22)), while in a non-inertial frame there appears, in addition, a pseudo-force in the equation as in (3-52). The latter does not depend on the effect of other bodies on the particle under consideration and, instead, is determined by the acceleration of the frame of reference. If the non-inertial frame does not possess a rotational motion relative to an inertial one, then the pseudo-force is given by (3-53). The question of identification of inertial frames in practice receives an answer in Newton’s first law.

A complete description of the motion of the particle is arrived at by solving the equation of motion, for which one needs a set of initial conditions, i.e., information relating to the state of motion of the particle at some chosen instant of time. Such a complete description is obtained relatively easily in the special case for rectilinear or planar motions

under a *constant force*, as we have seen above. More generally, the momentum and position of the particle can be determined in terms of its state of motion at the initial time by solving the equation of motion. The latter is a *differential equation of the second order* in time, since it can be written in the form

$$m \frac{d^2 \mathbf{r}}{dt^2} = \mathbf{F}(\mathbf{r}), \quad (3-81)$$

and needs two integrations for its solution (check eq. (3-81) out; in $\mathbf{a} = \frac{d\mathbf{v}}{dt}$, put $\mathbf{v} = \frac{d\mathbf{r}}{dt}$). Eq. (3-81) is the three dimensional generalization of eq. (3-36a).

In writing eq. (3-81), I have assumed that the force acting on the particle at time t depends on its position vector at that time, but not explicitly on t . Forces depending explicitly on t (referred to as time dependent forces) may also act on the particle, as in the case of forced simple harmonic motion (see sec. 4.6). Moreover, the expression for the force \mathbf{F} in the above equation will, in general, involve the position vectors of those particles that act upon it to produce the force. In the present context, these other position vectors are assumed to remain constant. More generally, however, the position vectors of all these other particles will also change with time, in which case one will have to solve the equations of motion of a *system* of particles (see sec. 3.17).

The constants of integration appearing in the process of solving the equation of motion (3-81) are determined in terms of the initial position and momentum of the particle (see, for instance, sec. 4.1.1). One can thereby determine, for a given force acting on the particle, the state of motion of the particle at any arbitrarily chosen instant of time in terms of its state of motion at the initial time instant.

The concepts of work and energy, introduced in sec. 3.15, are of fundamental importance in mechanics and, in particular, in the mechanics of a single particle. The change in the kinetic energy of the particle undergoing a motion under the action of a force is defined in terms of its initial and final velocities, and equals the work done on it by the force. In the case of a conservative force, this work is also equal to the difference between the initial and final potential energies, a result that implies the principle of

conservation of the total energy of the particle. This principle helps us obtain the first integral of the equation of motion of the particle, which constitutes a partial solution to the problem of determining the motion of the particle under a given force, for specified initial conditions, since the energy, which remains conserved, can be treated as one of the two constants of integration, and there remains one more independent constant to be determined in terms of the initial conditions, by means of one single step of integration of a first order differential equation..

While the concept of a particle is an idealization, the equations of motion for a particle and its solutions in various situations are of great practical relevance since these give us complete information about the motion of the *center of mass* (see sec. 3.17.3) of a system of particles in corresponding real life situations. In particular, one obtains the center of mass motion of *rigid bodies*.

While the concept of a rigid body is once again an idealized one, it happens to have considerable practical relevance since numerous bodies in real life can be described as rigid in an approximate sense.

For instance, if the body under consideration be a small rigid homogeneous sphere, then equation (3-81) may be invoked to obtain the motion of the center of the sphere. For some purposes, this constitutes a reasonably good description of the motion of the sphere as a whole.

3.17 Mechanics of a system of particles

3.17.1 Internal and external forces. Equations of motion.

I have talked so long mainly of the mechanics of a single particle. We shall now have a brief look at the mechanics of a *system* of particles.

In this, for the sake of convenience, let us think of a system of just two particles to start with, which we name A and B, their masses being respectively m_A and m_B (say).

The kinematic and dynamical quantities relating to the two particles will be denoted by suffices A and B respectively. Think of the particle A first. The force acting on this particle at any given instant can be thought to be made up of two distinct forces, of which one is the force exerted on it by B, while the other is the force exerted by particles other than B, i.e., by ones external to the system under consideration. We write the former of the two forces as \mathbf{F}_{AB} and call it an *internal* force relating to the system under consideration (note that, of the two suffices used here, the first suffix indicates the particle on which the force acts, while the second indicates the particle that exerts the force). The latter, the *external* force acting on A, will be denoted as \mathbf{F}_A . The total force acting on the particle A is then $\mathbf{F}_{AB} + \mathbf{F}_A$.

In a similar manner, the total force acting on B can be written as $\mathbf{F}_{BA} + \mathbf{F}_B$, of which the first term denotes the internal force and the second term the external force. One can now write out the equations of motion of the two particles in any inertial frame of reference as

$$\frac{d\mathbf{p}_A}{dt} = \mathbf{F}_{AB} + \mathbf{F}_A, \quad (3-82a)$$

$$\frac{d\mathbf{p}_B}{dt} = \mathbf{F}_{BA} + \mathbf{F}_B. \quad (3-82b)$$

Here $\mathbf{p}_A = m_A \mathbf{v}_A$ and $\mathbf{p}_B = m_B \mathbf{v}_B$ stand for the momenta of the two particles, and the expressions on the left hand sides represent their rates of change of momentum.

One can draw a number of important conclusions from these equations of motion relating to the system of particles. But before coming to these we shall first have to have a look at *Newton's third law* while, on the way, I will give you the definition of the *center of mass* of a system of particles.

3.17.2 Newton's third law

Newton's third law states, in effect, that the mutual forces of interaction between any pair of particles are equal and opposite to each other. Considering the pair made up

of the particles A and B referred to above, this means that the forces \mathbf{F}_{AB} and \mathbf{F}_{BA} are related to each other as

$$\mathbf{F}_{AB} = -\mathbf{F}_{BA}. \quad (3-83)$$

Two illustrations of the above relation can be found in 3-12(A) and (B). Notice that, even as both the figures conform to eq. (3-83), the forces \mathbf{F}_{AB} and \mathbf{F}_{BA} in 3-12 (A) act along the line connecting the two particles A and B, while in 3-12(B) they *do not* act along the connecting line. Forces of interaction of these two types are referred to as *central* and *non-central* forces respectively.

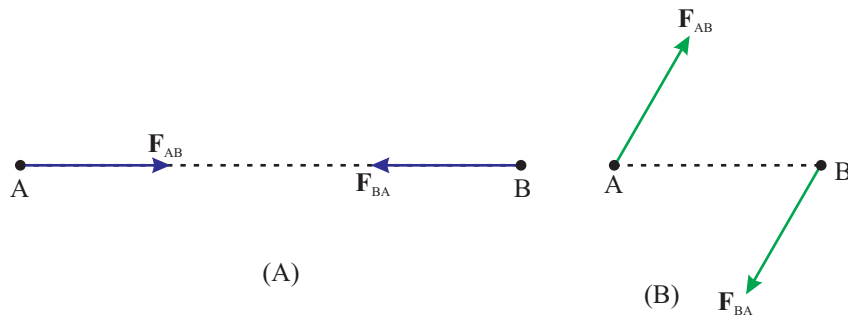


Figure 3-12: Illustration of Newton's third law, (A) central force, and (b) non-central force; in each case eq. (3-83) is satisfied.

Examples of interaction between particles by means of central force are the gravitational interaction, and the interaction between two charges in accordance with *Coulomb's law* of electrostatics. Non-central forces are seen to arise between a pair of ideal electrical dipoles or magnetic dipoles. The force between a pair of nucleons (the common name given to protons and neutrons) by means of what is known as *strong interaction* is known to involve a non-central force.

One can extend the statement of Newton's third law to two systems of particles, say, S_1 and S_2 . Each of the particles making up the first system experiences a force exerted by each particle belonging to the second system. Adding up all such forces, one obtains the total force, say, \mathbf{F}_1 , acting any specified instant of time, on S_1 due to S_2 . The force

\mathbf{F}_2 on S_2 due to S_1 is similarly obtained. Newton's third law then states that

$$\mathbf{F}_1 = -\mathbf{F}_2. \quad (3-84)$$

Calling the forces \mathbf{F}_1 and \mathbf{F}_2 the forces of action and reaction between the systems S_1 and S_2 , one can state Newton's third law in the more familiar form: *To every action there is an equal and opposite reaction.*

1. The terms action and reaction are more appropriately employed to describe the *impulses* exerted by the two systems on one another. Newton's third law can be formulated in terms of impulses as well (see sec. 3.17.8.1).
2. While defining the total force on either of the two systems S_1 and S_2 , one need not consider the internal forces exerted by particles within that system on one another because such forces cancel out in pairs when the total force on the system is worked out. Here we do not consider possible forces exerted by systems *other* than S_1 and S_2 . In the context of mutual interaction between S_1 and S_2 , such forces are to be treated as *external* ones.

3.17.3 Center of mass of a system of particles

3.17.3.1 The position and velocity of the center of mass

Consider a system of particles or a body made up of, say, N number of particles with masses, say, m_1, m_2, \dots, m_N . Let the position vectors of these particles with reference to any chosen origin, say, O at any given instant of time be respectively $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N$.

Imagine now a point, say, C , whose instantaneous position vector with respect to O is given by

$$\mathbf{R} = \frac{m_1\mathbf{r}_1 + m_2\mathbf{r}_2 + \dots + m_N\mathbf{r}_N}{m_1 + m_2 + \dots + m_N}. \quad (3-85)$$

Then this point is termed the *center of mass* (abbreviated as CM) of the system of particles under consideration at the given instant of time. This point may or may not correspond to the position of any particle belonging to the system of particles at the in-

stant of time considered. The center of mass bears a special relevance in the mechanics of a system of particles.

Think of a Cartesian co-ordinate system with O as the origin. Then the co-ordinates of the center of mass with reference to this co-ordinate system will be

$$\begin{aligned} X &= \frac{m_1x_1 + m_2x_2 + \dots + m_Nx_N}{m_1 + m_2 + \dots + m_N}, \\ Y &= \frac{m_1y_1 + m_2y_2 + \dots + m_Ny_N}{m_1 + m_2 + \dots + m_N}, \\ Z &= \frac{m_1z_1 + m_2z_2 + \dots + m_Nz_N}{m_1 + m_2 + \dots + m_N}. \end{aligned} \quad (3-86)$$

Here $(x_1, y_1, z_1), (x_2, y_2, z_2), \dots, (x_N, y_N, z_N)$ are the co-ordinates of the particles belonging to the system at the instant under consideration.

In the special case when all the particles constituting the system lie on a single straight line, one can choose the origin and the co-ordinate system in such a way that the said straight line coincides with the x-axis, as a result of which the y- and z co-ordinates of the center of mass become both zero, i.e., $Y = 0, Z = 0$ in (3-86). In other words, the center of mass of a system of particles all lying on a straight line, will also lie on that straight line. In a similar manner, the center of mass of a system of particles lying in a plane will also lie in that plane.

The following problems state important results relating to centers of mass of systems of particles.

Problem 3-19

Suppose that, for a system of particles all lying on the x-axis, the minimum and maximum values of the co-ordinates are a and b ($a \leq b$). Show that the co-ordinate (X) of the center of mass of the system will lie in the interval between a and b : $a \leq X \leq b$.

Answer to Problem 3-19

HINT: If x_i ($i = 1, \dots, N$) be the co-ordinates of the particles, then $a \leq x_i$ for all i , and hence

$$X = \frac{\sum m_i x_i}{\sum m_i} \geq \frac{\sum m_i a}{\sum m_i}, \text{ i.e., } a \leq X. \text{ Similarly, } X \leq b.$$

Problem 3-20

Suppose that three particles lying on the x-axis have their CM at the point X . Let the total mass of two of these be m and imagine that these two are replaced with a single particle of mass m located at the CM of these two. Show that the CM of this (imagined) particle and the remaining, third, particle will also be at X .

Answer to Problem 3-20

SOLUTION and NOTE: Let the masses of the particles be m_1, m_2, m_3 , where $m_1 + m_2 = m$, and their position co-ordinates be respectively x_1, x_2, x_3 , where the center of mass of the first two is at, say, X' , given by $X' = \frac{m_1 x_1 + m_2 x_2}{m_1 + m_2}$. The CM of mass m placed at X' and m_3 placed at x_3 will be at $\frac{mX' + m_3 x_3}{m + m_3}$, which is the same as $X = \frac{m_1 x_1 + m_2 x_2 + m_3 x_3}{m_1 + m_2 + m_3}$.

This shows that one can replace, in a system of particles, any subset of particles of total mass, say, m with a single particle of mass m located at the CM of that subset without altering the location of the CM of the whole system, i.e., the CM of this single particle of mass m and the subset of the remaining particles will be located at the same position as the CM of the system we started with.

It is often convenient to make use of this result in determining the CMs of various bodies and systems of particles.

The center of mass of a rigid body (i.e., a body all of whose constituent particles are at fixed distances from one another; this is an idealized but useful concept) is a point fixed rigidly in the body. It can be checked (try this out) that any rigid translation or rotation of the body results in the same translation or rotation of the center of mass.

Problem 3-21

Suppose that the instantaneous velocities of two particles of mass m_1 and m_2 are, respectively, v_1

and \mathbf{v}_2 . Show that the instantaneous velocity of their c.m. is

$$\mathbf{V} = \frac{m_1 \mathbf{v}_1 + m_2 \mathbf{v}_2}{m_1 + m_2}. \quad (3-87)$$

Answer to Problem 3-21

Take the time derivative of the instantaneous position vector of the center of mass $\mathbf{r} = \frac{m_1 \mathbf{r}_1 + m_2 \mathbf{r}_2}{m_1 + m_2}$.

The result stated in eq. (3-87) can be extended in a straightforward manner to give the velocity of the center of mass of a system of more than two particles:

$$\mathbf{V} = \frac{\sum_i m_i \mathbf{v}_i}{\sum_i m_i}, \quad (3-88)$$

where the notation is self-explanatory.

3.17.3.2 center of mass momentum

Referring to the result obtained in the last problem, assume that a particle of mass $M = m_1 + m_2$ is imagined to be located at the instantaneous position of the CM of the two particles and that this imagined particle keeps on moving with the CM as the latter changes its position along with the two particles under consideration. The motion of this imagined particle is then said to constitute the *center of mass motion* of the system. The momentum (\mathbf{P}) associated with this center of mass motion is, by (3-87),

$$\mathbf{P} = M\mathbf{V} = m_1 \mathbf{v}_1 + m_2 \mathbf{v}_2 = \mathbf{p}_1 + \mathbf{p}_2, \quad (3-89)$$

where \mathbf{p}_1 and \mathbf{p}_2 refer to the momenta of the two particles. This result extends to a system of more than two particles: the momentum associated with the center of mass motion (or, the center of mass momentum in brief) of a system of particles equals the vector sum of the momenta of the particles making up the system:

$$\mathbf{P} = \sum_i \mathbf{p}_i, \quad (3-90)$$

where, once again, the notation is self-explanatory.

3.17.3.3 Determination of the center of mass

The determination of the center of mass of bodies of various shapes is of considerable practical importance. The basic formula to use is eq. (3-85). In determining the center of mass of a body with a continuous mass distribution, it is often imagined to be made up of a large number of small mass elements, with a typical mass element of mass, say, δm being located at a point with co-ordinates (x, y, z) relative to any given co-ordinate system. In order to determine any of the three co-ordinates (X, Y, Z) of the center of mass (say, X) one has to sum up products of the form $\delta m x = \rho \delta v x$ (ρ being the density of the material at the point of location of the element under consideration) and then divide by M , the total mass of the body. In the limit of δm being made to be vanishingly small, the summation reduces to an integration over the entire body.

center of mass of a symmetric body

The determination of the position of the center of mass becomes relatively simple if the body is of a *symmetric* shape. For instance, if the body possesses a *reflection symmetry* about a given plane (say, P), then the center of mass has to lie in this plane.

Here the term 'reflection symmetry' does not only mean a symmetry of its geometrical shape, but relates to a symmetry of the mass distribution as well. For instance, consider a mass element δm of the body located at the point (x, y, z) . If there be an identical mass element belonging to the body at the point $(x, y, -z)$ for each and every mass element of the body, then the x-y plane will be a plane of reflection symmetry for it. Evidently, the z-co-ordinate of the center of mass of a pair of mass elements related by reflection symmetry about the x-y plane will be zero, and hence, the z-co-ordinate of the center of mass of the entire body will also be zero.

If, now, the body possesses two such planes of reflection symmetry, then the center of mass will lie on the line of intersection of the two planes (reason this out). Choosing the

x-axis of a co-ordinate system on this line, the co-ordinate of the center of mass will, as explained above, be given by a formula of the form $X = \frac{1}{M} \sum \delta m x$, where the summation reduces to an integration in the limit of the mass elements being vanishingly small.

In the special case of the body under consideration possessing two intersecting lines of symmetry, where a line of symmetry is the intersection of two planes of symmetry, the center of mass lies at the point of intersection of these two lines. For instance, the center of mass of a homogeneous spherical shell (or of a homogeneous filled sphere) is located at the center of the sphere. Similarly, the center of mass of a homogeneous cylinder (or a cylindrical shell) is located on its axis in the median plane (the 'center' of the cylinder).

The center of mass of a homogeneous hemispherical body

As an illustration of the way the center of mass of a body can be determined, consider a hemispherical body of mass, say, M and of radius R (see fig. 3-13), the mass distribution being assumed to be uniform. Choosing the origin of a co-ordinate system at the center of the sphere (of which the body under consideration constitutes a part) and the z-axis as shown in the figure, the hemispherical body can be imagined to be made up of a large number of thin discs, a typical disc, shown in the figure, being at a distance z above the center and having a thickness, say, δz .

The mass of the disc is then $\delta m = \frac{M}{\frac{2}{3}\pi R^3} \pi(R^2 - z^2)\delta z$ (reason this out; the radius of the disc is $\sqrt{(R^2 - z^2)}$). The co-ordinate of the center of mass, on the z-axis (an axis of symmetry for the hemispherical body) is then obtained by summing up expressions of the form $\delta m z$ and dividing by M , which reduces to

$$Z = \frac{1}{M} \int_0^R \frac{M}{\frac{2}{3}\pi R^3} \pi(R^2 - z^2) z dz = \frac{3}{8} R. \quad (3-91)$$

This, then, gives the height at which the center of mass is located above the center O of the hemisphere, on its axis OZ.

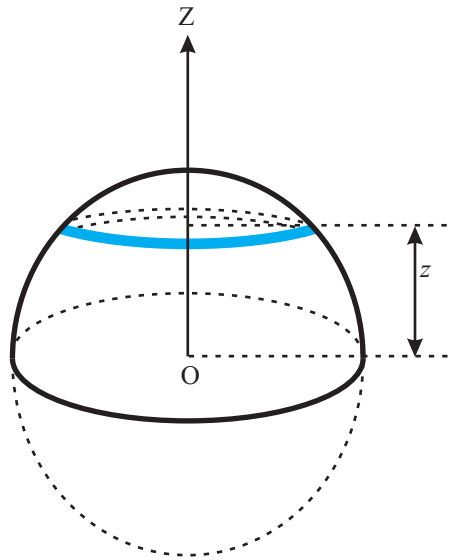


Figure 3-13: Illustrating the center of mass of a homogeneous hemispherical body of radius R ; the dotted line completes the sphere; the origin (O) of a co-ordinate system is chosen at the center of the sphere, and the z-axis is chosen to pass through the pole; a thin slice of width δz is considered at a height z above the equatorial plane; the center of mass lies on the z-axis at a height $\frac{3}{8}R$ above the center of the sphere.

3.17.4 System of particles: center of mass motion

Equations (3-82a), (3-82b) give the general forms of the equations of motion pertaining to a system made up of two particles A and B. We can draw a number of conclusions from these equations that can be generalized to a system made up of any number of particles.

Notice that these equations involve both internal and external forces where the former arise due to the interactions of the particles making up the system and the latter due to the effect of particles or bodies other than the particles in the system under consideration. As the particles move in accordance with these equations of motion, the position of their center of mass keeps on changing (in accordance with eq. (3-85)). Imagine that a particle of mass $M = m_A + m_B$ is located at the center of mass, moving along with the latter. As I have mentioned above, the motion of this imagined particle is referred to as the center of mass motion of the system, its momentum being

$$\mathbf{P} = \mathbf{p}_A + \mathbf{p}_B, \quad (3-92)$$

(see eq. (3-89) where suffices 1 and 2 have been used instead of A and B while referring to the two particles). This has been referred to above as the center of mass momentum of the system under consideration.

The left hand sides in equations (3-82a), (3-82b) represent the rates of change of the momenta of the two particles. Adding up the corresponding sides of these two one gets, on the left hand side, the rate of change of the center of mass momentum:

$$\frac{d\mathbf{P}}{dt} = (\mathbf{F}_{AB} + \mathbf{F}_{BA}) + (\mathbf{F}_1 + \mathbf{F}_2). \quad (3-93a)$$

In this equation, the first sum on the right hand side is zero in accordance with Newton's third law expressed in eq. (3-83), while the second sum represents the total external force on the system under consideration. Denoting this by the symbol \mathbf{F} , one has

$$\frac{d\mathbf{P}}{dt} = \mathbf{F}. \quad (3-93b)$$

Evidently, if there were a force \mathbf{F} acting on the imagined mass M located at the CM of the system, then the equation of motion of that particle would be precisely eq. (3-93b) (compare with eq. (3-22)). This statement applies to any system of particles in general: if the total mass of the particles making up a system be M and the vector sum of the external forces acting on the particles be \mathbf{F} , then the center of mass motion can be described as the motion of an imagined particle of mass M under the force \mathbf{F} .

More precisely, assuming that the imagined particle of mass M coincides with the center of mass at any chosen initial time instant and its velocity is the same as the center of mass velocity at that instant (as given by eq. (3-87) or (3-88)), its motion under the force \mathbf{F} tells us how the center of mass moves and where it is located at all other instants of time.

Notice that Newton's third law implies that the *internal forces of the system need not be considered in determining the center of mass motion*. As a corollary one concludes that, if the total external force acting on the particles of the system under consideration be zero

(i.e., if $\mathbf{F}_A = -\mathbf{F}_B$ in equations (3-82a), (3-82b)), then the rate of change of the center of mass momentum will be zero, i.e., in other words, the center of mass will move with uniform velocity. This statement applies to any system of particles in general: *if the total external force on the particles making up a system is zero, then the center of mass of the system moves with uniform velocity.*

If one thinks of an inertial system with respect to which the center of mass is at rest at any chosen initial instant of time, then the center of mass will continue to remain at rest in that frame. This is then referred to as the *center of mass frame* of the system. Thus, for instance, if the origin in the center of mass frame is chosen to coincide with the center of mass then the position vectors of the two particles at any instant of time will be related as

$$\mathbf{r}_B = -\frac{m_A}{m_B}\mathbf{r}_A. \quad (3-94a)$$

If, instead of a pair of particles, one considers two systems of particles or bodies A and B, and if there be no external force acting on the composite system, then in the center of mass frame one again will have an equation of the form (3-94a) where now \mathbf{r}_A , \mathbf{r}_B refer to the centers of mass of the bodies A and B and where it is assumed that the center of mass coincides with the origin at any chosen instant of time (thereby remaining coincident with it at all times). In particular, if the trajectory of one of the centers of mass (say, that of A) remains confined to, say, the x-axis then that of the other (i.e., the center of mass of B) will also remain confined to the x-axis, the relation between the two co-ordinates being given by

$$x_B = -\frac{m_A}{m_B}x_A. \quad (3-94b)$$

3.17.5 Principle of conservation of momentum

If the total external force acting on the particles of a system is zero then in any inertial frame of reference the center of mass of the system moves with uniform velocity (V , given by eq. (3-88)) and hence the center mass momentum ($\mathbf{P} = M\mathbf{V}$; also referred to

as, simply, the momentum of the system) remains unchanged with time. This is known as the *principle of conservation of momentum*.

In the special case of the frame of reference being the center of mass frame, the constant value of the center of mass momentum (i.e., the total momentum of a system of particles) reduces to zero.

In the relativistic theory, the center of mass frame is *defined* as the one in which the total momentum of a system of particles is zero.

3.17.6 System of particles: principle of conservation of energy

The concepts of kinetic and potential energies explained for a single particle can be extended to a system of particles as well. If the masses of the particles making up the system be m_1, m_2, \dots, m_N and if their instantaneous velocities be respectively $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N$, then the kinetic energy of the system is given by the expression

$$K = \frac{1}{2}m_1\mathbf{v}_1^2 + m_2\mathbf{v}_2^2 + \dots + m_N\mathbf{v}_N^2. \quad (3-95)$$

The system possesses an ability to perform work on other systems to the extent of this amount by virtue of motion of its constituent particles.

The potential energy of the system can be defined if the motion of each of its constituent particles occurs under a *conservative* field of force. As I have mentioned, the particles are acted upon by internal as also by external forces. All these forces will be said to form a conservative system if the following condition is satisfied.

Suppose that the system under consideration is taken from an initial reference configuration to any other configuration, with each particle following a path of its own, where a configuration corresponds to a set of positions of the constituents particles of the system. The same change of configuration can be brought about with the particles following *other* sets of paths as well. For each of these various sets of paths one can calculate the work done by the internal and external forces on the particles of the systems as they

follow their own paths in the change of configuration. If for all these different ways by which the change in configuration of the system of particles is brought about, the work done by the internal and external forces be the same, say, W then, the force system will be said to be conservative one.

More generally, one need not even consider the paths followed in possible motions of the system in changing over from one configuration to another. The condition for the system of forces to be conservative demands that the sum of the line integrals of the forces acting on all the particles considered severally along the individual paths followed by these particles along *any arbitrary* set of such paths (corresponding to the given initial and final configurations of the system), not necessarily the paths followed in possible motions of the system, be independent of these paths chosen.

Assuming that the above condition for the force system to be conservative is satisfied, the potential energy of the system of particles in the final configuration with reference to the chosen standard configuration is defined as the total work W in a change of configuration along any possible path, taken with a negative sign:

$$V = -W. \quad (3-96)$$

The system of particles under consideration possesses an ability to perform this amount of work on other systems by virtue of its configuration, i.e., the positions of its constituent particles, relative to the reference configuration. In other words, the system can be made to perform this amount of work on other systems if brought back from the final configuration to the reference configuration in an appropriate manner.

The principle of conservation of energy, explained earlier in the context of the motion of a single particle, can be extended to a system of particles moving under a conservative system of internal and external forces: the total energy, i.e., the sum of kinetic and potential energies of the system remains conserved during any given motion of the system.

The potential energy of a system of particles, defined as above, can be, in general, decomposed into two parts - the *mutual potential energy of the particles of the system* and the potential energy of the particles under the external forces acting on these.

3.17.7 Elastic and inelastic collisions

3.17.7.1 Introduction

Fig. 3-14 shows a particle P approaching another particle Q from a large distance along the straight line AB with a velocity v_1 , where the particle Q is assumed, for the sake of simplicity, to be initially at rest (more generally, Q may be moving initially with a uniform velocity v_2 , but we assume for the time being that $v_2 = 0$). As long as P remains at a large distance from Q, the forces of interaction between the two may be ignored, and P moves like a free particle along the straight line AB with uniform velocity v_1 .

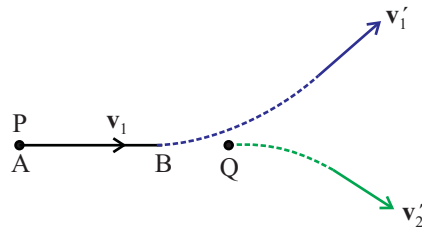


Figure 3-14: Depicting a collision; P and Q are the two colliding particles, of which Q is assumed to be initially at rest in the frame of reference chosen (the 'laboratory frame'; more generally, Q may have an initial velocity v_2); P approaches Q with an initial velocity v_1 ; after the collision, the two particles have velocities v'_1 , v'_2 .

As P approaches to within a small distance of Q, interaction forces between the two are brought into play and the trajectory of P (dotted line in the figure) deviates from a straight line. Eventually, as the two particles get separated from each other by a large distance, they move like two free particles with velocities v'_1 and v'_2 . Such an event where two particles, initially at a large distance from each other, interact on coming close together and then get separated so as to move effectively like free particles again, is termed a *collision*. Our description of the collision process assumes an inertial frame of reference.

3.17.7.2 Elastic and inelastic processes

A collision is, in general, a complex process. Imagine, for instance, that P and Q are two rubber balls, each having a finite size. Each of the two balls may then possess a *rotational* motion in addition to the translational motion of its center of mass and the process of collision may affect the distribution of the kinetic energy of the ball between the rotational and translational modes. In the present context, we assume that the bodies under consideration are so small that their rotational kinetic energies may be ignored, allowing us to consider these as point particles coincident with the respective centers of mass.

The rotational kinetic energies may be negligible owing to the relevant moments of inertia being small.

However, even when the rotational energies are negligibly small, there may remain other structural aspects of the colliding particles that one cannot possibly ignore. For instance, the two rubber balls may get *deformed* as they collide and a part of their translational kinetic energy may be transformed into elastic energy of deformation. Part of this energy ultimately gets *dissipated* into the materials of the two balls as heat or, more precisely, their *internal energy* (see sec. 8.6).

This requires that, along with the translational kinetic energies of the particles, their internal energies (i.e., energies associated with their internal constituents) be taken into consideration in the energy accounting for the process. If the inter-conversion of energy between the translational modes and the internal modes can be ignored and the energy accounting of the process can be adequately done in terms of the translational kinetic energies of the particles (or bodies) alone, then the collision is said to be an *elastic* one. If, on the other hand, the exchange of energy with the internal modes cannot be ignored, the collision is *inelastic* in nature.

The energy associated with internal modes of a particle or body need not always manifest itself as internal energy in the thermodynamic sense (see chapter 8 for an introduction

to a number of basic concepts in thermodynamics, including that of internal energy). For instance, if a composite particle (a system having a negligible volume but still having an internal structure of its own) is made up of a relatively small number of constituent particles bound to one another, it may relate to the *binding energy* of these constituent parts. Thus, the constituents may become less strongly bound if some part of kinetic energy gets converted into internal energy of the system. It may even be possible for one or more bound constituents to get dissociated from the body and move away as one or more independent particles. This is precisely what happens in a number of reactions involving nuclei and elementary particles.

Thus, elastic collisions are idealized processes where the energies associated with internal modes of the colliding particles or bodies can be ignored in the energy accounting of these processes, while real collision processes are, in general, inelastic. In an inelastic collision, it is possible that the identities of the particles P and Q may get altered and one may end up with a new set of particles, say P', Q', ..., after the collision. For the present, however, we assume that the particles retain their identity as they move away from each other like free particles with velocities $\mathbf{v}'_1, \mathbf{v}'_2$.

3.17.7.3 The energy balance equation

When P and Q are at a large distance from each other, their potential energy of interaction can be neglected, and their total energy can be expressed as the sum of their kinetic energies of translation (where we ignore their rotational energies) and their internal energies. Thus, considering two such configurations with P and Q separated from each other by a large distance, one before and the other after the collision process (these are referred to as the *initial* and the *final* configurations respectively), the *principle of conservation of energy* gives

$$\frac{1}{2}m_1\mathbf{v}_1^2 + \frac{1}{2}m_2\mathbf{v}_2^2 + Q_1 = \frac{1}{2}m_1\mathbf{v}'_1{}^2 + \frac{1}{2}m_2\mathbf{v}'_2{}^2 + Q_2. \quad (3-97)$$

In this equation Q_1 and Q_2 denote the total energy associated with the internal modes of the particles in the initial and final configurations. We have, moreover, taken into consideration the kinetic energy of the particle Q in the initial configuration, which was

assumed to be zero in sec. 3.17.7.1 (and in the illustration in fig. 3-14) for the sake of simplicity.

We write this equation in the more compact form

$$K = K' + Q, \quad (3-98)$$

where K and K' are the total kinetic energies in the initial and final configurations, and $Q \equiv Q_2 - Q_1$ stands for the energy converted into energy of the internal modes.

Elastic collisions then correspond to $Q = 0$, while the more general case of inelastic collisions corresponds to $Q \neq 0$. We shall, moreover, consider only those inelastic collisions for which $Q > 0$.

1. Inelastic collisions with $Q < 0$ are observed in nuclear processes especially when there occurs a change in identity of the colliding particles like, for instance, in a process of the form $P+Q \rightarrow P'+Q'$. Such processes are referred to as *exoergic* ones. By contrast, commonly observed collisions are characterized by $Q > 0$ where some kinetic energy is converted into internal energy of the colliding bodies rather than the other way round. For instance, as a rubber ball hits and rebounds from a wall, it gets deformed at the moment of impact, and part of the deformation energy is eventually dissipated as heat into the material of the ball. Another instance of an inelastic collision where there occurs a decrease in the internal energy is found in *Raman scattering* where a *photon* (an energy quantum of electromagnetic radiation, see chapter 16) collides with a molecule and, under appropriate conditions, comes off with a *higher energy*, the additional energy coming at the expense of the internal energy of the molecule. Collision events among two particles with $Q < 0$ are sometimes referred to as *superelastic* ones.
2. Including the rotational kinetic energies in the energy accounting does not alter the above considerations relating to elastic and inelastic collisions. The rotational energies go into the expressions for K and K' , while not altering the definition of Q_1 , Q_2 , and Q .
3. In a nuclear reaction, a certain quantity termed the *Q-value* is defined so as to define quantitatively the extent to which internal binding energy is converted into

kinetic energy in the reaction. In the context of such a reaction, the quantity Q , as defined in eq. (3-98), happens to be numerically the same as its Q -value, but with a *negative* sign.

3.17.7.4 Momentum balance

Regardless of whether the collision under consideration is an elastic or an inelastic one, the total *momentum* of the system made up of P and Q in the initial configuration has to equal the total momentum in the final configuration, as required by the *principle of conservation of momentum*. Indeed, whatever processes take place involving P and Q, along with their internal constituents, these are all due to forces of an *internal* nature in relation to the composite system made up of P and Q so that, in the absence of external forces (arising due to possible interactions with some *other* bodies) acting on these two, the total linear momentum of P and Q in the initial configuration has to be the same as that in the final configuration (see sec. 3.17.5).

The equation expressing the principle of conservation of momentum in the present context reads

$$m_1 \mathbf{v}_1 + m_2 \mathbf{v}_2 = m_1 \mathbf{v}'_1 + m_2 \mathbf{v}'_2, \quad (3-99)$$

or, equivalently, in an obvious notation,

$$\mathbf{p}_1 + \mathbf{p}_2 = \mathbf{p}'_1 + \mathbf{p}'_2, \quad (3-100)$$

3.17.7.5 Relative velocities: normal and tangential

In a realistic description of a collision process, P and Q are at times required to be treated as extended bodies. In numerous situations of practical importance, one can assume these to be rigid bodies that come into contact at a point (or at points lying on a surface) momentarily, and then separate from each other (see fig. 3-15). The *relative velocity* of P with respect to Q just before the contact can then be resolved into a component *normal* to the common surface of contact (\mathbf{v}_n) and one tangential to this

surface (v_t). Considering the two components (v'_n , v'_t) of the relative velocity just after the collision, one often finds that the tangential component remains unchanged in the collision process ($v_t = v'_t$) while the normal components are related as

$$v'_n = -e v_n, \quad (3-101)$$

i.e., the normal component gets reversed in direction and is reduced by a factor e in magnitude, where e depends on details relating to the two bodies involved in the collision process.

The interaction between the two bodies occurs by impact, at the time of which an *impulsive force* (refer to sec. 3.17.8) acts on these. Assuming that there is no force in the nature of friction (see sec. 3.23) between the bodies, the impulse acts in a direction normal to the surface of contact at the time of collision, and the change of momentum of either body occurs along this direction.

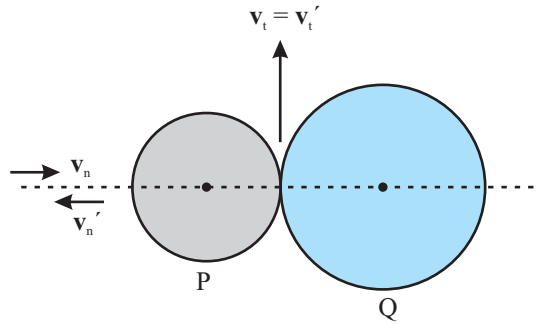


Figure 3-15: Illustrating the normal and tangential components of relative velocity for two rigid spheres coming in contact in a collision process; v_n and v_t are the normal and tangential components before the collision, while v'_n , v'_t are the corresponding components after the collision.

This coefficient e is referred to as the *coefficient of restitution* characterizing the collision. An elastic collision corresponds to $e = 1$ (as we shall see, $e = 1$ implies $Q = 0$ in eq. (3-98)) while a collision with $e = 0$ is termed a *perfectly inelastic* one.

Since the coefficient of restitution is defined in terms of *relative velocities*, its definition does not depend on the frame of reference used. In particular, the relation (3-101) holds in the *center of mass frame* (see sec. 3.17.7.6) of the two colliding bodies.

Collisions observed in our daily experience or in engineering practice often involve an impact by contact. In these cases, the characterization of a collision in terms of the coefficient of restitution is often convenient. For collisions involving microscopic particles, on the other hand, a characterization in terms of the quantity Q introduced in eq. (3-98) instead of the coefficient of restitution, is usually more meaningful.

3.17.7.6 The center of mass frame

In sec. 3.17.7.1 I mentioned that our description of the collision process holds in any inertial frame of reference. In an inertial frame, the center of mass of a closed system (in the present instance, the system made of the particles P and Q), i.e., one not acted upon by any external force, moves with a constant velocity, which is another way of stating the principle of conservation of momentum.

If, then, we consider a frame of reference in which the center of mass of the system is *at rest*, then that system will also be an inertial one. This frame of reference, termed the *center of mass frame* (see sec. 3.17.4) is characterized by the center of mass momentum being zero at every stage of the collision process and hence, in particular, in the initial and the final configurations. In other words, both sides of eq. (3-99) are zero in the center of mass frame.

This makes the description of the collision process especially simple in the center of mass frame. Thus, of the two final velocities \mathbf{v}'_1 and \mathbf{v}'_2 , only one (say, \mathbf{v}'_1) remains undetermined since the other is determined from the condition $m_1\mathbf{v}'_1 + m_2\mathbf{v}'_2 = 0$ (the use of *momenta* in the place of velocities makes the description even more simple; see problem 3-23). The magnitude of this velocity can then be determined by taking into consideration other principles relating to the collision like, for instance, the one expressed by the relation (3-101).

3.17.7.7 Describing the collision process

However, eq. (3-101) does not express a fundamental principle relating to the collision process. It is in the nature of an observed feature of a class of collision processes,

namely those that can be described as *impacts* of rigid bodies on one another. As mentioned above, an impact is a collision process where the interaction between the colliding bodies is brought into play only at the instant at which they come into actual contact with each other, when they experience an *impulsive* force (see sec. 3.17.8). This is evidently an idealization since, more generally a collision involves deformable bodies and, moreover, the interaction between the colliding bodies does not require actual contact.

As I have said, a collision process is, in general, a complex one and there does not exist a single unified approach for the description of all kinds of collision processes. In physics, one commonly encounters collision processes between extended bodies that can be assumed to have various idealized properties on the one hand, and those involving microscopic particles or groups of particles on the other. The former are the collisions we observe in our everyday experience and those involved in engineering theory and practice. The latter, on the other hand, are the ones relevant in the domain of atomic and molecular physics. An attempt to describe both of these two types of collision processes in a single theoretical and descriptive framework may lead to confusions.

For instance, the collisions between microscopic bodies can often be described as those involving point particles while an attempt to reduce a collision between extended bodies to that between point particles may lead to inconsistencies. In a collision between microscopic particles, one needs to consider the detailed interaction forces (or potentials) between the particles so as to obtain a complete description of the collision process, and a formulation in terms of an impact does not have much use.

3.17.7.8 'Head-on' collision

In the case of two rigid bodies, a 'head-on' collision is a collision by impact for which, in the center of mass frame, the tangential component of the relative velocity is zero both before and after the collision. In this case, the condition of zero tangential velocity implies that the two colliding bodies approach each other along a straight line and, after the impact, recede from each other along the same straight line. Such a simple description holds for a head-on collision between two particles as well. On writing down

the energy and momentum balance equations in the center of mass frame, one finds that the total kinetic energies in the initial and final configurations are related to each other (see problem 3-23) as

$$K' = e^2 K. \quad (3-102)$$

where e stands for the coefficient of restitution as defined in eq. (3-101). In other words, under the conditions assumed, a fraction e^2 of the initial kinetic energy in the center of mass frame is retained in the final configuration, the rest being converted into energy associated with internal modes of the colliding bodies. In particular, $e = 1$ implies that there is no conversion of kinetic energy into internal modes (elastic collision) while a value $e = 0$ implies that the entire kinetic energy in the center of mass frame is converted into internal energy, as result of which the final configuration consists of the two colliding bodies sticking together, both at rest in the center of mass frame (perfectly inelastic collision). In any other inertial frame the two bodies, while sticking together in the final configuration, possess a common non-zero velocity.

Thus, a head-on collision can be described, in the center of mass frame, as a collision between two point particles for which all the initial and final velocities are along a single straight line, say, the x-axis of a co-ordinate system, for which the following relations hold good:

$$\frac{1}{2}(m_1 v_1'^2 + m_2 v_2'^2) = \frac{1}{2}e^2(m_1 v_1^2 + m_2 v_2^2), \quad (3-103a)$$

$$m_1 v_1 + m_2 v_2 = m_1 v_1' + m_2 v_2' = 0, \quad (3-103b)$$

$$v_2' - v_1' = -e(v_2 - v_1), \quad (3-103c)$$

where v_1, v_2, v_1', v_2' stand for the x-components of the respective vectors or, in other words, the magnitudes, with appropriate *signs* of these vectors. It may be mentioned that, in view of the relations (3-103b), eq. (3-103a) follows from (3-103c) (check this out;

refer to problem 3-23 below).

Knowing $v_1, v_2 (= -\frac{m_1}{m_2}v_1)$, one can then determine $v'_1, v'_2 (= -\frac{m_1}{m_2}v'_1)$ from eq. (3-103c) (or, equivalently, eq. (3-103a)). One thereby arrives at the following result

$$v'_1 = -ev_1, \quad v'_2 = -ev_2. \quad (3-104)$$

This shows that one indeed arrives at simple relations when one refers to the center of mass frame. Let the center of mass frame, in which the final velocities are related to the initial velocities through the simple-looking relations (3-104), be C.

3.17.7.9 Head-on collision in the 'laboratory frame'

The same head-on collision can also be looked at from any other inertial frame (say, S) as well. Consider, in particular, an inertial frame whose velocity with respect to the center of mass frame lies along the line of collision so that, in this frame also, the collision is one dimensional, i.e., in this frame the two colliding particles move along the same line both before and after the collision.

More generally, in an arbitrarily chosen inertial frame the collision is no longer head-on, and the motion of the colliding particles turns out to be a planar one.

Suppose that the velocity of the center of mass frame C with respect to the frame S is u . In the special case in which $u = -v_2$, the particle Q has an initial velocity zero in S (here v_2 represents the initial velocity of Q in C). This corresponds to a common experimental situation in which a particle P (the 'projectile') is projected towards a particle Q (the 'target'), where the latter is initially stationary in the laboratory. In this instance, then, S stands for the 'laboratory frame', in which the projectile P has the velocity $u_1 = v_1 + u = (1 + \frac{m_1}{m_2})v_1$. Conversely, if the initial velocity of P in the laboratory frame be u_1 , then its velocity in the center of mass frame will be $v_1 = \frac{m_2}{m_1 + m_2}u_1$.

Starting from the relations (3-104), one can work out the expressions relating the initial

and final velocities (u_1, u_2, u'_1, u'_2) in the frame S, which one finds to be

$$u'_1 = \frac{m_1 u_1 + m_2 u_2 + e m_2 (u_2 - u_1)}{m_1 + m_2}, \quad u'_2 = \frac{m_1 u_1 + m_2 u_2 + e m_1 (u_1 - u_2)}{m_1 + m_2}. \quad (3-105)$$

Problem 3-22

Check the above relations out.

Answer to Problem 3-22

HINT: The velocities in the frame S are related to the corresponding velocities in the center of mass frame C as

$$u_1 = v_1 + u, \quad u_2 = v_2 + u, \quad u'_1 = v'_1 + u, \quad u'_2 = v'_2 + u.$$

Now use the relations (3-103b) and (3-104).

For an elastic head-on collision between two particles of equal mass, it follows from the above relations that $u'_1 = u_2, u'_2 = u_1$, i.e., the particles *exchange their velocities*.

In other words, in a head-on elastic collision between particles of equal mass, the two particles simply swap their energies. If, however, the particles are of unequal mass, then their energies get changed.

Problem 3-23

Consider a head-on collision between two particles P, Q, of masses m_1, m_2 . If p be the momentum of P in the center of mass frame before the collision, and e be the coefficient of restitution, find the kinetic energies of the two particles before and after the collision (a) in the center of mass frame, and (b) in the laboratory frame, in which Q is initially at rest. Hence verify eq. (3-102), and write down the corresponding formula in the laboratory frame.

Answer to Problem 3-23

HINT: (a) In the center of mass frame, the initial momenta of P and Q are, respectively, p , $-p$, and the velocity of P relative to Q is $v_1 = \frac{p}{m_1} + \frac{p}{m_2}$. Hence the momenta after the collision are $p' = -ep$, $-p'$ (reason this out). Thus, the kinetic energies before and after the collision are, respectively, $\frac{p^2}{2m_1}, \frac{e^2 p^2}{2m_1}$ (for P) and $\frac{p^2}{2m_2}, \frac{e^2 p^2}{2m_2}$ (for Q). This gives $K = \frac{p^2}{2}(\frac{1}{m_1} + \frac{1}{m_2})$, and $K' = e^2 K$ (verifying eq. (3-102)). (b) Since the velocity of Q before the collision is $-\frac{p}{m_2}$ in the center of mass frame and 0 in the laboratory frame, all velocities in the laboratory frame are obtained from the corresponding velocities in the center of mass frame by adding $\frac{p}{m_2}$ to the latter. Thus, the velocities of P, Q in the laboratory frame after the collision are $-\frac{ep}{m_1} + \frac{p}{m_2}$, and $\frac{ep}{m_2} + \frac{p}{m_2}$. Hence the kinetic energies before the collision are $\frac{p^2}{2m_1 m_2^2}(m_1 + m_2)^2$ and 0, while those after the collision are $\frac{p^2}{2m_1}(\frac{m_1}{m_2} - e)^2$ and $\frac{p^2}{2m_2}(1 + e)^2$. This gives, in the laboratory frame,

$$K' = K \frac{m_1 + e^2 m_2}{m_1 + m_2}, \quad (3-106)$$

Note that this reduces to $K' = K$ for an elastic collision ($e = 1$) while, in the limit $m_1 \rightarrow 0$, it gives $K' = e^2 K$, as it should (reason out why).

3.17.7.10 Elastic collisions in planar motion

The interactions of molecules in an ideal gas are conveniently described as elastic collisions. Elastic collisions are also relevant in describing the molecular interactions in a real gas. Numerous other types of processes in the microscopic domain are also in the nature of elastic collisions.

A head-on collision considered in sec. 3.17.7.8 is one where the motion of the colliding particles (recall that, in the case of a collision involving extended bodies, the latter can, at times, be represented by point particles) remains confined to a straight line. More commonly, however, collisions between particles (as also between extended bodies) involve motion in two or three dimensions.

Fig. 3-16 depicts a collision between two particles in planar motion in the center of mass frame, the latter being characterized by the fact that the momenta of the two particles are equal and opposite at every instant throughout the collision process. Even so, the two momenta may not be directed along the straight line joining the instantaneous

positions of the particles at every instant of time. The distance between the lines of motion of the particles in the initial configuration is referred to as the *impact parameter* characterizing the collision.

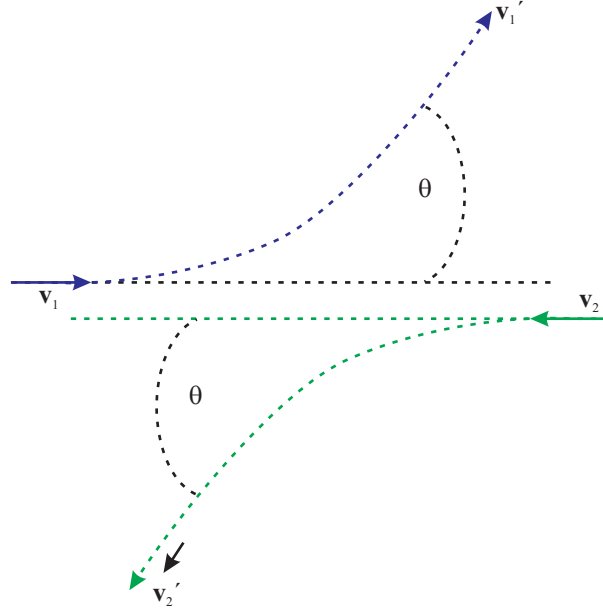


Figure 3-16: Depicting an elastic collision between two particles in planar motion; in the center of mass frame, the initial velocity vectors \mathbf{v}_1 , \mathbf{v}_2 get rotated by an angle θ so as to give the final velocity vectors \mathbf{v}_1' , \mathbf{v}_2' ; the distance between the initial lines of motion of the two particles is the impact parameter.

If the force of interaction between two particles be central in nature (see sec. 3.17.2) then a planar motion of the two particles in an inertial frame under their mutual interaction is possible. For non-central interactions, on the other hand, planar motion is, in general, not possible. In practice, however, non-central forces are relevant only in the sub-atomic domain.

Since the collision is elastic, the energy balance equation is of the form (3-103a) with $e = 1$,

$$\frac{1}{2}(m_1\mathbf{v}_1 \cdot \mathbf{v}_1 + m_2\mathbf{v}_2 \cdot \mathbf{v}_2) = \frac{1}{2}(m_1\mathbf{v}_1' \cdot \mathbf{v}_1' + m_2\mathbf{v}_2' \cdot \mathbf{v}_2'), \quad (3-107)$$

in which we explicitly acknowledge that the velocities are vectors where, in the present

instance, they all lie in a plane.

When this is supplemented with the momentum equation valid in the center of mass frame,

$$m_1 \mathbf{v}_1 + m_2 \mathbf{v}_2 = 0, \quad m_1 \mathbf{v}'_1 + m_2 \mathbf{v}'_2 = 0, \quad (3-108)$$

it is found that the respective velocities in the final configuration, i.e., after the collision, are equal in magnitude to the corresponding velocities in the initial configuration:

$$|\mathbf{v}'_1| = |\mathbf{v}_1|, \quad |\mathbf{v}'_2| = |\mathbf{v}_2|, \quad (3-109)$$

(check this out).

This implies that the two directed lines containing the vectors \mathbf{v}'_1 , \mathbf{v}'_2 are obtained by simply rotating the directed lines containing \mathbf{v}_1 , \mathbf{v}_2 about an axis perpendicular to the plane of collision by a certain angle θ , referred to as the *scattering angle* in the center of mass frame.

1. In the domain of processes involving microscopic particles, a collision is often referred to as *scattering*.
2. A planar collision for which the impact parameter is zero effectively corresponds to a head-on collision where the scattering angle is $\theta = \pi$. More generally, the scattering angle θ lies in the range $0 \leq \theta \leq \pi$.

Thus, the description of an elastic collision in planar motion is quite simple in the center of mass frame (C): the velocity vectors get rotated through an angle θ , with their magnitudes remaining unchanged.

Consider now any other inertial frame S with respect to which the velocity of the center of mass frame C is, say, \mathbf{u} . In this frame, the velocities of the particles before and after

the collision are related to those in the center of mass frame as

$$\mathbf{u}_1 = \mathbf{v}_1 + \mathbf{u}, \mathbf{u}_2 = \mathbf{v}_2 + \mathbf{u}, \mathbf{u}'_1 = \mathbf{v}'_1 + \mathbf{u}, \mathbf{u}'_2 = \mathbf{v}'_2 + \mathbf{u}. \quad (3-110)$$

If the velocity \mathbf{u} of the center of mass frame C relative to S (the velocity of S relative to C is equal and opposite) is chosen to lie in the plane of motion in the center of mass frame, then the motion remains a planar one when described in the frame S . In particular, one can choose an inertial frame in which one of the particles, say, Q is initially at rest (i.e., $\mathbf{u}_2 = 0$). This describes a situation in which a projectile P is scattered by a target particle Q where the latter is initially stationary in the frame under consideration. In this case, S corresponds to the so-called 'laboratory frame', in which the motion remains confined to the plane containing the initial position of the target and the initial direction of motion of the projectile.

In addition, the *relative motion* of the two particles remains confined to a plane.

Looking at the collision in the frame S (assuming that the motion is a planar one in S) and making use of the momentum and the energy balance equations, one can relate the final velocities in S to the initial velocities, where now these relations involve an additional parameter that can be chosen as θ , the scattering angle in the center of mass frame (an alternative choice for this additional parameter is the impact parameter b ; the relation between the two parameters b and θ is determined by the nature of interaction between the particles). In this case it is found that even when the particles are of the same mass, the elastic collision does not simply swap the energies of the two particles. One can work out the amount of energy transfer between the two particles, where one finds that energy may get transferred from the more energetic particle to the less energetic one or, in certain circumstances, even in the reverse direction (refer to section 3.17.7.11 below for more detailed considerations; the total energy, however, remains constant).

Problem 3-24

Consider an elastic collision between two particles P and Q, of mass 0.03 kg and 0.02 kg respectively where, in the laboratory frame, Q is at rest and P approaches along the x-axis of a rectangular co-ordinate system with a velocity $2 \text{ m}\cdot\text{s}^{-1}$. In the center of mass frame, the line of motion of P gets rotated by $\frac{\pi}{3}$ towards the y-axis, the motion being confined in the x-y plane. What is the angle that the line of motion of Q after the collision makes with the x-axis in the laboratory frame?

Answer to Problem 3-24

The velocity of the center of mass frame relative to the lab frame is (all units in SI system implied) $V = \frac{2 \times 0.03}{0.05} = 1.2$, and the velocities of P and Q in the CM frame before the collision are 0.8 and -1.2 respectively along the x-axis. The final velocity of P in the CM frame is thus 0.8 along a direction making an angle $\frac{\pi}{3}$ with the x-axis, and that of Q is 1.2 along a direction making an angle $\frac{-2\pi}{3}$ with the x-axis (obtained by rotating through an angle $\frac{2\pi}{3}$ from the x-axis in the clockwise sense). In the lab frame, then, the velocity components of Q are $v_x = 1.2 + 1.2 \cos \frac{-2\pi}{3} = 0.6$ and $v_y = 1.2 \sin \frac{-2\pi}{3} = -0.6 \times \sqrt{3}$. The angle made by the direction of motion with the x-axis is $\theta = \tan^{-1} \frac{v_y}{v_x} = \tan^{-1}(-\sqrt{3}) = -\frac{\pi}{3}$.

3.17.7.11 Direction of energy transfer in elastic collisions

Fig. 3-17 depicts a planar collision between two particles in any chosen frame of reference in which a Cartesian co-ordinate system is chosen such that the motion of the particles is confined to the x-y plane. A convenient description of the collision is obtained by referring to the velocity of the center of mass (this remains constant throughout the collision), say, v , making an angle ϕ with the x-axis (we choose $v > 0$ without loss of generality), and the velocity of either particle relative to the center of mass. We label the two particles with indices '1' and '2', and choose the scale of mass such that $m_1 = 1$, and $m_2 = \mu$, for the sake of convenience. If the velocity, relative to the center of mass, of particle '1' before the collision be u , making an angle, say, ψ with the direction of the center of mass velocity (again, we choose $u > 0$ without loss of generality), then the corresponding velocity for particle '2' will be $\frac{u}{\mu}$ in the opposite direction (this makes their total momentum zero in the center of mass frame).

Thus, the velocity components of the two particles (along the x-and y-axes of the co-

ordinate system in the chosen frame) before the collision are:

$$\begin{aligned} u_{1x} &= u \cos \psi + v \cos \phi, \quad u_{1y} = u \sin \psi + v \sin \phi \\ u_{2x} &= -\frac{u}{\mu} \cos \psi + v \cos \phi, \quad u_{2y} = -\frac{u}{\mu} \sin \psi + v \sin \phi. \end{aligned} \quad (3-111)$$

It now remains to specify the velocity components after the collision, which is conveniently accomplished by specifying the angle of rotation of the velocity vector of either particle in the center of mass frame (refer to sec. 3.17.7.10). Denoting the angle of rotation by θ , the velocity components after the collision are seen to be

$$\begin{aligned} v_{1x} &= u \cos(\psi + \theta) + v \cos \phi, \quad v_{1y} = u \sin(\psi + \theta) + v \sin \phi \\ v_{2x} &= -\frac{u}{\mu} \cos(\psi + \theta) + v \cos \phi, \quad v_{2y} = -\frac{u}{\mu} \sin(\psi + \theta) + v \sin \phi. \end{aligned} \quad (3-112)$$

Thus, the same collision can be described in terms of two alternative sets of parameters specifying the velocities before collision: $\{u, v, \phi, \psi\}$ and $\{u_{1x}, u_{1y}, u_{2x}, u_{2y}\}$; the two sets are related as in (3-111). Since the collision is an elastic one, only one additional parameter, namely, θ , is needed to specify the velocity components after the collision.

The energies of the two particles, as calculated in the chosen frame of reference, are then

$$\begin{aligned} \text{(before collision)} \quad E_1 &= \frac{1}{2}(u^2 + v^2 + 2uv \cos \alpha), \quad E_2 = \frac{1}{2}\left(\frac{u^2}{\mu} + \mu v^2 - 2uv \cos \alpha\right) \\ \text{(after collision)} \quad E'_1 &= \frac{1}{2}(u^2 + v^2 + 2uv \cos(\alpha + \theta)), \quad E'_2 = \frac{1}{2}\left(\frac{u^2}{\mu} + \mu v^2 - 2uv \cos(\alpha + \theta)\right) \quad (\alpha \equiv \psi - \phi), \end{aligned} \quad (3-113)$$

We now restrict our considerations to a collision in which $E_1 > E_2$, i.e., the first particle has a higher energy before the collision. In terms of the parameters u, v, ϕ, ψ this means that

$$(E_1 > E_2 \Rightarrow) \cos \alpha > \frac{1}{4uv} \left(u^2 \left(\frac{1}{\mu} - 1 \right) + v^2 (\mu - 1) \right), \quad (3-114)$$

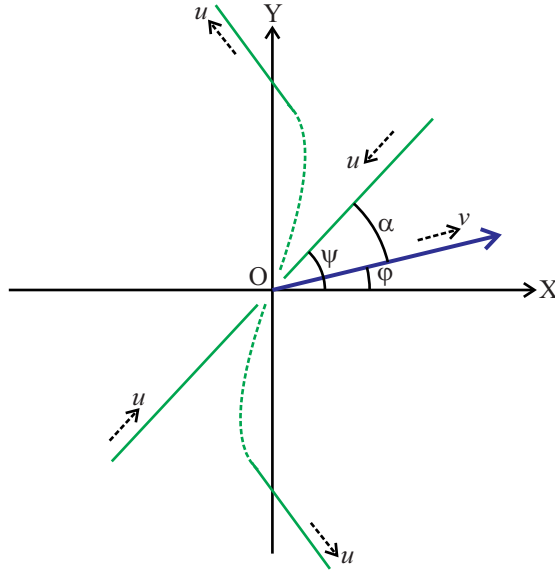


Figure 3-17: Elastic collision between two particles in a chosen frame of reference, and in any chosen plane, which is taken to be the x-y plane in a Cartesian co-ordinate system; the velocity of the center of mass in this frame is v , making an angle ϕ with the x-axis, while the velocity of one of the two particles (marked '1', the other particle is marked '2') relative to the center of mass is u making an angle α with the direction of velocity of the latter, before the collision (according to our choice of the scale of mass, this is also the momentum of either particle in the center of mass frame); as a result of the collision, the velocity of particle '1' relative to the center of mass gets rotated by an angle θ (the velocity of particle '2' suffers a rotation in the opposite direction); on performing an average over possible values of θ and α , one finds that the more energetic of the two particles loses energy in the collision.

where, as defined above, $\alpha = \psi - \phi$.

Subject to this condition, we will now examine whether $E'_1 - E_1$ is positive or negative, where (3-113) gives

$$E'_1 - E_1 = uv(\cos(\alpha + \theta) - \cos \alpha). \quad (3-115)$$

Recall that we have chosen both u and v to be positive, leaving ϕ and ψ unspecified, which means that $\alpha = \psi - \phi$ can have any arbitrary value subject to the inequality (3-114). If, now, the collision be such that the right hand side of (3-115) is positive, then that would mean that energy is *gained* by the particle that posses a higher energy compared to the other before the collision. On the other hand, in the case that the right hand side of (3-115) is less than zero, the more energetic of the two particles (particle '1')

in the present context) *loses* energy in the collision. Evidently, α and θ can be chosen such that either of these possibilities is realized, i.e., in other words, it is possible for the more energetic of the two particles to either gain or lose energy in the collision.

The question now arises as to whether, considering all possible values of α (subject to (3-114); note that a necessary condition for this inequality to hold is that the right hand side is to be less than unity) and θ , the faster of the two particles can be said to lose energy *on the average*. For this, note that if the right hand side of the inequality (3-114) is less than unity, then there exists an angle α_0 in the range $0 \leq \alpha_0 \leq \pi$ such that $-\alpha_0 \leq \alpha \leq \alpha_0$. Thus, the averaging is to be performed with α uniformly distributed over this range, and θ uniformly distributed in the range $0 \leq \theta \leq \pi$ (see sec. 3.17.7.10). On averaging first over the scattering angle θ (this is equivalent to averaging over all possible impact parameters), one obtains

$$\langle E'_1 - E_1 \rangle_{\theta} = -uv \left(\frac{\sin \alpha}{\pi} - \cos \alpha \right), \quad (3-116a)$$

where the suffix θ denotes an averaging over the scattering angle. Finally, averaging over all the allowed values of α , one obtains

$$\langle E'_1 - E_1 \rangle_{\theta, \alpha} = -uv \left(\frac{\sin \alpha_0}{\alpha_0} \right), \quad (3-116b)$$

which means that, *on the average, the faster of the two particles loses energy* in an elastic collision.

Here we have considered, for the sake of simplicity, collisions confined to some specified plane in any chosen frame of reference. If we now consider all such possible planes, the conclusion reached above continues to hold.

3.17.8 Impulse. Impulsive forces.

3.17.8.1 Impulse of a force or of a system of forces

Consider a force \mathbf{F} acting on a particle. In addition to being a function of position, denoted by the position vector \mathbf{r} , the force may depend explicitly on time (t) as well. For

instance, the force may act on the particle for a short duration of time, say, from an initial time t_0 to $t_0 + \tau$, where τ is a small time interval, whereas, outside this interval, the force is zero.

The dependence of \mathbf{F} on t may be indicated by writing \mathbf{F} as a function of t , i.e., in the form $\mathbf{F}(t)$.

The *integral* of $\mathbf{F}(t)$ over time, is referred to as the *impulse* (\mathbf{I}) of the force. Assuming that the force is non-zero only in an interval t_0 to $t_0 + \tau$, the impulse is given by the expression

$$\mathbf{I} = \int_{t_0}^{t_0 + \tau} \mathbf{F}(t) dt, \quad (3-117)$$

where, more generally, the lower and the upper limits of integration are to be taken as $-\infty$ and $+\infty$ respectively.

Making use of the equation of motion of the particle (eq. (3-22)), one can relate the impulse of the force acting on a particle to its change in momentum:

$$\mathbf{I} = \int \frac{d\mathbf{p}}{dt} dt = \mathbf{p}_2 - \mathbf{p}_1, \quad (3-118)$$

where the lower and upper limits of integration are left implied, and where \mathbf{p}_1 and \mathbf{p}_2 are the momenta of the particle at the initial and final instants (i.e., at t_0 and $t_0 + \tau$ respectively or, more generally, at times $t \rightarrow -\infty$ and $t \rightarrow \infty$).

This expression of the impulse in terms of the change of momentum of the particle under consideration, at times makes it a more useful quantity to work with than the force itself, especially when a large force acts for a very short duration, as a result of which its actual value as a function of time cannot be known with precision. In such a situation, the impulse is a more well defined quantity than the force, and is more convenient to use since it relates directly to the change in the state of motion of the

particle rather than to the *rate* of change.

Considering two particles, say A and B, in mutual interaction, with no external force acting on these, we have already seen that the forces \mathbf{F}_{AB} and \mathbf{F}_{BA} are equal and opposite to each other (see sec. 3.17.2), a result which holds at every instant of time. Consequently, the impulse on A due to the force (\mathbf{F}_{AB}) exerted by B will be equal and opposite to that on B due to the force (\mathbf{F}_{BA}) exerted by A:

$$\mathbf{I}_A = \int \mathbf{F}_{AB}(t)dt = - \int \mathbf{F}_{BA}(t)dt = -\mathbf{I}_B. \quad (3-119)$$

Thus, the terms ‘action’ and ‘reaction’ in Newton’s third law (see sec. 3.17.2) can be interpreted as referring to the impulses exerted by the two particles or bodies (as indicated below, the concept of impulse may be extended to a system of particles as well) on each other. Used along with eq. (3-118), this immediately leads to a re-derivation of the principle of conservation of momentum in the present context:

$$\mathbf{p}_{A2} - \mathbf{p}_{A1} = \mathbf{p}_{B1} - \mathbf{p}_{B2}, \text{ i.e., } \mathbf{p}_{A1} + \mathbf{p}_{B1} = \mathbf{p}_{A2} + \mathbf{p}_{B2}, \quad (3-120)$$

where the notation is self-explanatory.

One can, in a similar manner, define the impulse on a system of particles or a rigid body as

$$\mathbf{I} = \sum_i \int \mathbf{F}_i^{(\text{ext})}(t)dt = \int \mathbf{F}^{\text{ext}}(t)dt = \mathbf{P}_2 - \mathbf{P}_1, \quad (3-121)$$

where only the *external* forces ($\mathbf{F}_i^{(\text{ext})}$ ($i = 1, 2, \dots$)) on the particles making up the system are relevant since, by Newton’s third law, the internal forces (and hence their impulses) cancel in pairs. Further, in the above relation, \mathbf{P}_1 and \mathbf{P}_2 denote the initial and final *center of mass momenta* of the system under consideration since the total external force on

the system equals the rate of change of its center of mass momentum (see eq. (3-93b)).

This is an interesting and useful result: *the total impulse of the external forces acting on a system of particles over any given interval of time equals the change of its center of mass momentum in that interval.*

Considering two systems of particles (say, S_1 and S_2), the mutual impulses between them are equal and opposite, which implies the conservation of the total center of mass momentum of the composite system made up of S_1 and S_2 provided that there are no forces acting on these due to any *other* system.

As mentioned above, the impulse of a force is a useful concept in practice when the force acting on a particle or a body is in the nature of an *impulsive* one (see sec. 3.17.8.2).

3.17.8.2 Impulsive forces

Fig. 3-18(A) depicts schematically the time variation of a force F assumed, for the sake of simplicity, to act along a given straight line (say, the x-axis of a co-ordinate system; thus, F is represented by a scalar with a sign). According to this figure, the force has a value F_0 during an interval of time from $t = 0$ to $t = \tau$ (say), while it is zero for $t < 0$ and $t > \tau$. The impulse of the force is then $F_0\tau$, the area under the graph showing the variation of the force.

While in fig. 3-18(A), the force is shown to increase and decrease abruptly at the instants $t = 0$ and $t = \tau$ respectively, fig. 3-18(B) depicts the variation of force under more realistic conditions, where the force is seen to operate for a short time, but its increase and decrease occurs smoothly, though rapidly. The impulse of the force (in the present context, a scalar with a sign), being defined as the integral

$$I = \int F(t)dt, \quad (3-122)$$

is again represented by the area under the graph in the figure. In the following I refer to an idealized variation of the type (A) in fig. 3-18, for the sake of simplicity.

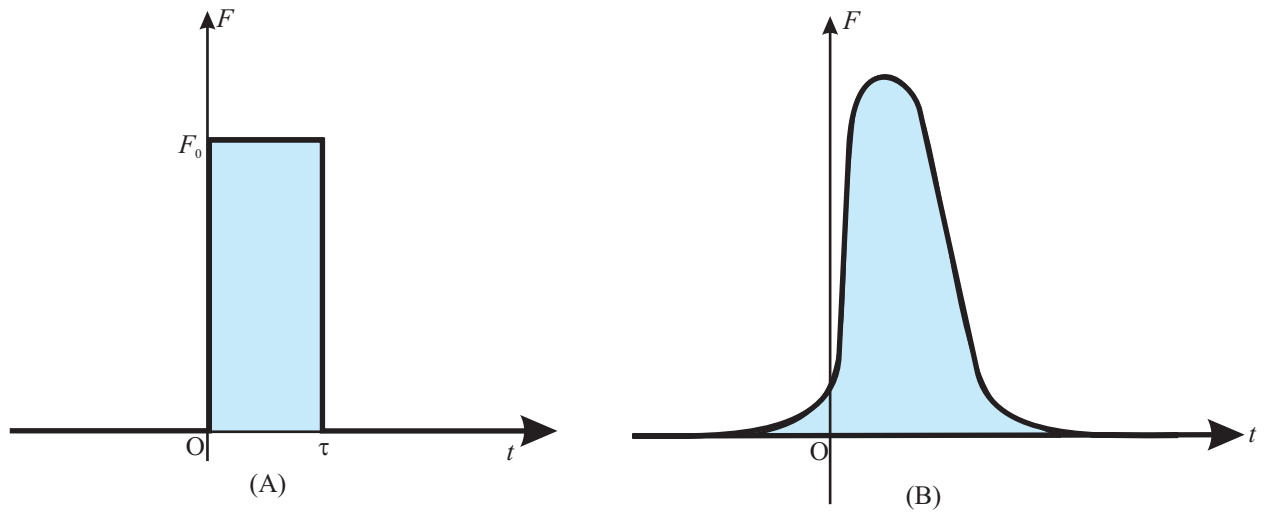


Figure 3-18: Graphical representation of an impulsive force: (A) step-function variation, where the force remains constant during the interval $t = 0$ to $t = \tau$, being zero outside this interval; (B) smooth variation where the force varies continuously, but is of a small magnitude outside a certain small interval of time.

Now suppose that the interval τ is a small one while, at the same time, F_0 is large so that the product $F_0\tau$ is of an appreciable magnitude. Such a force is said to be an *impulsive* one because, though short-lived, it is of a sufficient magnitude so as to cause an appreciable change of momentum in a particle or a body.

If the time of duration of the force τ is assumed to be small without F_0 being correspondingly large, the impulse $F_0\tau$ will be so small as to cause a negligible change in the momentum of the particle or the body on which the force acts.

The forces of action and reaction brought into play during the *impact* of two rigid (or approximately rigid) bodies are generally of an impulsive nature because such an impact lasts for only a short time while, at the same time, producing an appreciable change of momentum of each of the bodies under consideration.

Considering, for instance, an impact between two bodies (say, P and Q, see sec. 3.17.7.5) of masses m_1, m_2 , let \mathbf{u}_{1n} and \mathbf{u}'_{1n} be the normal components (i.e., components along the common normal to the two surfaces in contact when the impact takes place) of the ve-

locities of P before and after the impact respectively, the corresponding components of the relative velocity of P with respect to Q being \mathbf{u}_n and \mathbf{u}'_n . We assume that all these velocities are with respect to any given inertial frame of reference (say S). Let the corresponding velocities in the center of mass frame of the two bodies be denoted by \mathbf{v}_{1n} , \mathbf{v}'_{1n} , \mathbf{v}_n , \mathbf{v}'_n .

The change in the normal component of velocity of the particle P in the center of mass frame due to the impact is related to the normal component of the relative velocity of the particles before the impact as

$$\mathbf{v}'_n - \mathbf{v}_n = -\frac{m_2}{m_1 + m_2}(1 + e)(\mathbf{v}_{1n} - \mathbf{v}'_{1n}), \quad (3-123)$$

where e stands for the coefficient of restitution introduced in sec. 3.17.7.5.

Problem 3-25

Check this relation out.

Answer to Problem 3-25

Use eq. (3-101), and the relations $\mathbf{v}_{1n} = \frac{m_2}{m_1 + m_2}\mathbf{v}_n$, $\mathbf{v}'_{1n} = \frac{m_2}{m_1 + m_2}\mathbf{v}'_n$, which hold by the definition of the center of mass frame.

Describing the impact with reference to any other inertial frame, one then obtains a corresponding equation relating the change in the normal component of the velocity of the particle P in that frame to the normal component of the relative velocity before the impact. Correspondingly, the *impulse* of the force of impact on the body P, which may be assumed to be an impulsive one, is given by

$$\mathbf{I} = -\frac{m_1 m_2}{m_1 + m_2}(1 + e)\mathbf{u}_n, \quad (3-124)$$

where \mathbf{u}_n denotes the normal component of the relative velocity, in the inertial frame

under consideration, of P with respect to Q before the impact. The impulse on Q is then $-I$, assuming that the impact is not affected by any *other* body.

Problem 3-26

Check eq. (3-124) out.

Answer to Problem 3-26

Use the fact (see sec. 3.9.2) that the difference of any two velocities, whether taken at the same or at two different time instants, remains unchanged in the transformation between two inertial frames.

Problem 3-27

A bat hits a ball of mass 0.2 kg where the latter travels horizontally along the negative direction of the x -axis of a Cartesian co-ordinate system with a speed of $20 \text{ m}\cdot\text{s}^{-1}$ before being hit. After the impact, the ball moves with a speed of $30 \text{ m}\cdot\text{s}^{-1}$ making an angle $\frac{\pi}{3}$ with the positive direction of the x -axis towards the vertically upward y -axis, the motion being confined to the x - y plane. Calculate the impulse of the bat on the ball. Assuming that the force exerted by the bat on the ball acts for an interval of 0.1 s , and is constant during that interval, calculate the force.

Answer to Problem 3-27

The change in the momentum of the ball has components (all SI units implied) $\delta p_x = 0.2 \times (30 \cos \frac{\pi}{3} - (-20))$ along the x -axis and $\delta p_y = 30 \sin \frac{\pi}{3}$ along the y -axis. These are the components of the impulse of the bat on the ball. The force components are respectively $F_x = \frac{\delta p_x}{0.1}$ and $F_y = \frac{\delta p_y}{0.1}$.

3.18 Newton's laws and action-at-a-distance

In writing the equations of motion describing a system of particles, we have included for the sake of generality, the external forces on the system. These external forces can be assumed to be produced by the action of one or more systems of particles on the system under consideration. One can, in principle, consider then a bigger system where

all these other systems of particles are included, arriving, in the end, at a *closed* system of particles which is not acted upon by any external forces.

Considering for simplicity a closed system consisting of just two particles, the equations of motion pertaining to that system will be of the form of equations (3-82a), (3-82b) with the external forces F_A , F_B absent. Looking at the equation for any one of the two particles, one then finds that its rate of change of momentum at any given time t is equal to the force exerted on it by the other particle, where the expression of the force involves, in general, the position vectors of the two particles *at the same instant of time* t . It is this feature that makes the equations a set of second order differential equations in time.

What this means is that, for instance, the force exerted on the particle A at time t by the particle B is determined by the position of B at the instant t itself. This, in turn, means that the influence of B on A at time t is exerted *instantaneously* from its position at the same instant t . This evidently implies a very special assumption regarding the way the influence is transmitted from B to A. Such instantaneous action of one particle on another, implicit in the Newtonian equations of motion, is referred to as *action-at-a-distance*.

In reality, whatever influence B exerts on A, has to travel down in the form of some *signal* from B to A, and hence the influence felt by A at time t has to originate at some earlier time, i.e., has to depend on the position of B at that earlier instant. On the face of it, this renders the entire framework of Newtonian mechanics invalid.

Yet, the Newtonian equations do give a reasonably good description of motions of systems within a broad domain of experience. This, precisely, is the domain of *non-relativistic* mechanics where the speeds of the particles whose motions are described by the equations are all *small* compared to the speed of the signals causing the particles to exert forces on one another. Typically, the signals travel with the *speed of light* (c) and, for particle velocities small compared to c , one can, to a good degree of approximation, assume the signal velocity to be *infinitely large*. It is in this approximation that the

assumption of instantaneous transmission of the influence of one particle on another, and hence the principle of action-at-distance, can be assumed to be a valid one.

In a more complete theory, one has to take into account the *fields* that carry the influence of one particle on another in the form of *waves* of various descriptions, where these waves constitute the signals mentioned above. For instance, the electromagnetic field, made up of space- and time dependent electric and magnetic field strengths, is responsible for the interactions between charged particles (see chapter 14 for an introduction to electromagnetic waves). A theoretical framework based on the assumption of such field mediated action of one particle on another can be considered to be of a more fundamental nature compared to the Newtonian one. The reason why such a theory is to be considered to be a more fundamental one is that it does not need the artificial assumption of action-at-a-distance.

3.19 Angular motion

3.19.1 Angular velocity of a particle about a point

In fig. 3-19, O is any chosen origin while P represents the instantaneous position of a particle. Suppose that T stands for a small interval of time where, after this interval, the particle is located at P' . Then, for sufficiently small T , the instantaneous velocity of the particle has to be along PP' . The plane of the diagram has been chosen here to contain the instantaneous position and velocity vectors, i.e., the points O and P, and the line along which the particle moves at the given instant of time. In the interval T , the particle describes an angle $\angle POP'$ about O. This angle, which we call θ , will be a small one for sufficiently small T , and the ratio $\frac{\theta}{T}$ will give the instantaneous rate at which the particle describes angle about O.

More precisely, the limiting value $\lim_{T \rightarrow 0} \frac{\theta}{T}$ is referred to as the *angular velocity* of the particle at the position P about O. Denoting this as ω , one can define an *angular velocity vector* as follows. Imagine a unit normal vector \hat{n} perpendicular to the plane POP' , whose

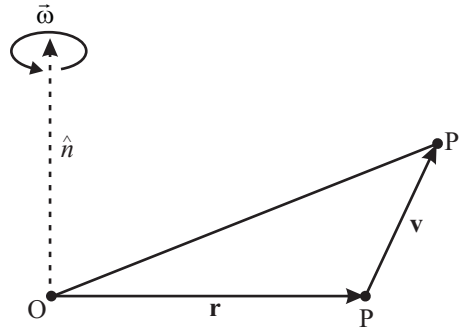


Figure 3-19: Illustrating the concept of angular velocity about a point; the position vector of the particles describes the angle $\angle POP' (= \theta)$ about O in a small time interval T ; the instantaneous rate of describing the angle is given by the ratio $\omega = \frac{\theta}{T}$ for $T \rightarrow 0$; the angular velocity vector $\vec{\omega}$ is directed along the perpendicular to the plane containing the position vector \mathbf{r} and the instantaneous velocity \mathbf{v} , being related to these two in a right handed sense.

sense is related to the directions of the instantaneous position vector (\mathbf{r}) and the velocity vector (\mathbf{v}) by the right hand rule.

There can be two unit vectors normal to the plane POP' , of which the one in the direction of the vector product $\mathbf{r} \times \mathbf{v}$ is related to \mathbf{r} and \mathbf{v} by the right hand rule. Indeed, the vector product is *defined* in the first place by the right hand rule: if a right handed cork-screw is rotated from \mathbf{r} to \mathbf{v} through the smaller of the two angles between the two vectors, then the direction in which the screw advances is defined to be the direction related to \mathbf{r} and \mathbf{v} by the right hand rule. An alternative (and more common) way to describe this is to imagine the thumb, index finger, and the middle finger of the right hand to be extended in such a manner that the middle finger is perpendicular to the plane containing the thumb and the index finger. The direction in which the middle finger extends is then said to be related to the directions of the thumb and the index finger by the right hand rule. You will find the right hand rule described in sec. 2.8.

One can then define a *vector angle* described about O as a vector of magnitude θ , directed along \hat{n} . The instantaneous rate of describing the vector angle by P about O gives the

instantaneous angular velocity vector ($\vec{\omega}$) of the particle about O in the position P:

$$\vec{\omega} = \frac{\theta}{T} \hat{n} = \omega \hat{n}, \quad (3-125)$$

where the limit $T \rightarrow 0$ is implied. In the following, we shall use the term angular velocity to denote either the scalar ω or the vector $\vec{\omega}$, depending on the context.

The angular velocity is given in terms of the position vector \mathbf{r} and velocity vector \mathbf{v} as

$$\vec{\omega} = \frac{\mathbf{r} \times \mathbf{v}}{r^2}. \quad (3-126)$$

Problem 3-28

Establish the relation (3-126) .

Answer to Problem 3-28

HINT: Imagine an arc of a circle drawn with O as center and with radius OP, in the plane OPP'; the length of the arc intercepted between OP and OP', in the limit of vanishingly small T , is $v \sin \theta T$, and the rate at which the position vector describes angle about O is $\frac{v \sin \theta}{r}$.

Note that there is a sense of rotation associated with a vector angle and an angular velocity vector, as shown, for instance, in fig. 3-19.

The direction and magnitude of the angular velocity about any given point O depends on the position and the instantaneous velocity of the particle and, in general, keeps on changing as the particle moves along its trajectory. At the same time, the angular velocity depends on the choice of the origin O. If, at any given instant, the vectors \mathbf{r} and \mathbf{v} are perpendicular to each other then the instantaneous motion of the particle about O is said to be one of *rotation*. In this case eq. (3-126) assumes the simpler form

$$\mathbf{v} = \vec{\omega} \times \mathbf{r}, \quad (3-127)$$

(check this out).

The unit of angular velocity in the SI system is *radian per second* ($\text{rad}\cdot\text{s}^{-1}$).

3.19.2 Angular velocity about an axis

In fig. 3-20, AB is any given straight line (which we term the *axis* in the present context) and P, P' are two positions of a particle at a small interval of time T . If perpendiculars PN and P'N' are dropped from P and P' respectively, on AB, then it is seen from the figure that the perpendicular dropped from the instantaneous position of the particle on the axis suffers a rotation about the latter as also a translation along it, in the course of time. The planes containing AB and PN and also AB and P'N' have been shown in the figure. If the angle between these two planes be θ , then the instantaneous rate of rotation about the axis AB is given by $\frac{\theta}{T}$ (or, more precisely, by its limiting value for $T \rightarrow 0$).

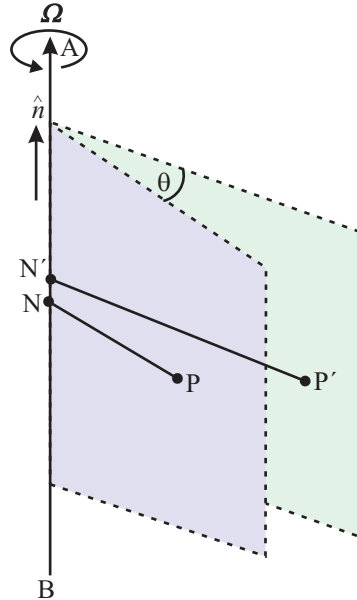


Figure 3-20: Illustrating the concept of angular velocity of a particle about an axis (AB); P and P' denote instantaneous positions of the particle at two instants separated by a small time interval; dropping normals (PN, P'N') on the axis, the angular velocity is the rate at which the normal turns about AB; the angular velocity *vector* is directed along \hat{n} , being related to the sense of rotation by the right hand rule.

This defines the angular velocity Ω of the particle about AB. If the unit vector along AB related to the sense of rotation of the particle by the right hand rule be denoted by \hat{n} , then the vector $\Omega\hat{n}$ is referred to as the *angular velocity vector* of the particle about the axis AB. Evidently, the magnitude and direction of the angular velocity depends on the choice of the axis. If the instantaneous motion of the particle be such that the rate of displacement of the foot of the perpendicular dropped on the axis from the instantaneous position of the particle is zero, then the motion is referred to as a *rotation* about the axis. Evidently, in a rotational motion, the instantaneous velocity of the particle will be directed along the perpendicular to the plane containing the axis AB and the line PN. The motion is, in that case, a rotation about the point N as well.

Problem 3-29

Referring to fig. 3-20, show that the angular velocity about an axis is given by the expression

$$\vec{\Omega} = \frac{(\vec{\rho} \times \mathbf{v}) \cdot \hat{n}}{\rho^2} \hat{n}, \quad (3-128)$$

where $\vec{\rho}$ stands for the vector extending from N to P, and \hat{n} is a unit vector along the axis, chosen along either of the two possible directions, not necessarily related to the sense of rotation by the right hand rule. If the direction of $\vec{\rho} \times \mathbf{v}$ makes an acute angle with \hat{n} then $\vec{\Omega}$ points along \hat{n} . If, on the other hand, the angle is obtuse, then it points in an opposite direction.

Answer to Problem 3-29

HINT: This expression is obtained by first looking at the angular velocity about the point N and then taking the component along the axis.

The relation between the angular velocity ω (eq. (3-126)) about a point O chosen on the axis AB and the angular velocity Ω about AB, is seen to be

$$\omega \cos \phi = \Omega \sin^2 \beta, \quad (3-129a)$$

where ϕ is the angle between $\vec{\omega}$ and \hat{n} (here \hat{n} is taken to be related to the sense of rotation by the right hand rule), and β that between \mathbf{r} (i.e., the position vector of the

particle relative to the origin O) and \hat{n} (fig. 3-21), where the latter may be acute or obtuse. Note that, in fig. 3-21, the dotted line through O (denoting the direction of ω), the axis AB, and the line OP are not, in general, coplanar.

Problem 3-30

Check eq. (3-129a) out.

Answer to Problem 3-30

HINT: Let $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$, where \mathbf{v}_1 is perpendicular to the plane containing the axis AB and the instantaneous position P of the particle (i.e., the plane containing \hat{n} and ρ), while \mathbf{v}_2 lies in this plane. Then

$$\Omega = \frac{v_1}{\rho}, \text{ and } \vec{\omega} \cdot \hat{n} = \omega \cos \phi = \frac{v_1}{r} \sin \beta.$$

In establishing the last relation, decompose \mathbf{v}_2 into two components, one along \hat{n} and another along $\vec{\rho}$.

In the special case where the instantaneous motion of the particle is one of rotation about AB (in this case the motion is one of rotation about O as well; however, the converse is not necessarily true), one has $\phi = \pm(\frac{\pi}{2} - \beta)$, depending on whether β is acute or obtuse, and

$$\omega = \Omega \cos \phi. \quad (3-129b)$$

This, however, excludes the case when the instantaneous motion is in the plane of \mathbf{r} and \hat{n} , in which case formula (3-129b), which is derived by canceling $\cos \phi$ from both sides of (3-129a), does not hold (since $\cos \phi = 0$ in this case).

3.19.3 Circular motion

Suppose that a particle moves along the circumference of a circle as in fig. 3-22, the center of the circle being, say, O. It is then said to execute a circular motion. In the figure, P denotes the instantaneous position of the particle and P' its position after a

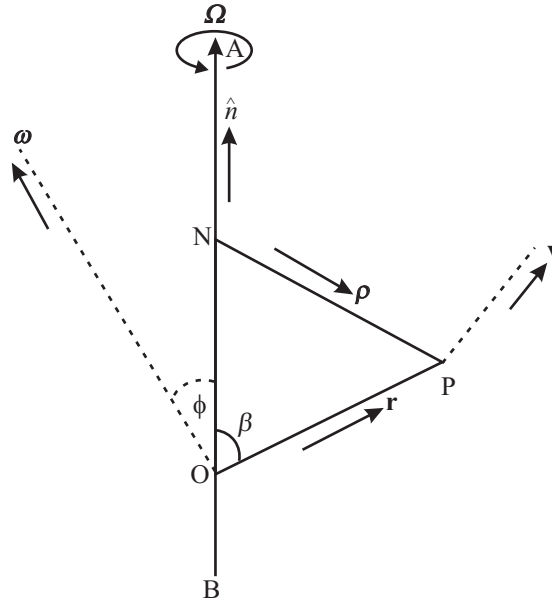


Figure 3-21: Illustrating the relation between angular velocity of a particle, located at P, about an axis (AB) and that about a point O chosen to lie on the axis (see sec. 3.19.2 for details).

small time interval, say, T . If the angle $\angle POP'$ be θ , then the distance traversed by the particle in time T is $a\theta$, where a stands for the radius of the circle. Then, for sufficiently small T (mathematically, in the limit $T \rightarrow 0$), the instantaneous angular velocity of the particle is $\omega = \frac{\theta}{T}$, while the instantaneous speed is $v = \frac{a\theta}{T}$. One therefore has the following relation between instantaneous speed and angular velocity

$$v = a\omega. \quad (3-130)$$

The direction of the velocity is along the tangent to the circle at P while that of the angular velocity vector is along the perpendicular to the plane of the circle related to the direction of motion by the right hand rule.

You can check that the above statements are consistent with eq. (3-127) for any given position of the particle on the circle. Indeed, a circular motion is a rotational motion about the center O for all positions of the particle on the circle. Moreover, considering the straight line perpendicular to the plane of the circle, the motion of the particle is

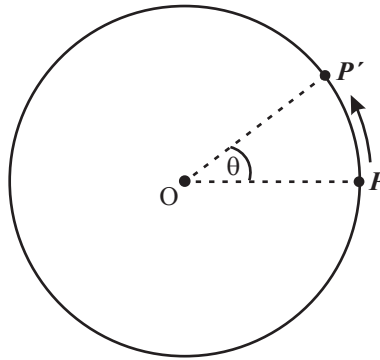


Figure 3-22: Illustrating circular motion of a particle; starting from the position P, the particle reaches the position P' in time T , describing an angle θ about the center O; the instantaneous speed and angular velocity are related by formula (3-130); the motion is one of rotation about O at every instant, as also a rotation about an axis passing through O and perpendicular to the plane of the circle.

also a rotation about that line as axis, with the same angular velocity ω .

If the angular velocity in a circular motion remains unchanged with time then the speed of the particle also remains constant. This is referred to as *uniform circular motion*. In this special case it is no longer necessary to refer to an infinitesimally small time interval in the definition of ω and v . Considering, for instance an arbitrary time interval t , the angle described about the center of the circle (referred to as the angular displacement) in this interval will then be given by

$$\theta = \omega t. \quad (3-131)$$

In other words, the relation between angular displacement, angular velocity, and time interval is analogous to the corresponding relation between displacement, velocity and time in uniform motion along a straight line.

In general, however, the angular velocity in a circular motion does not necessarily remain constant in time, which means that the angular *acceleration* need not be zero. The definition of angular acceleration in circular motion is analogous to that of acceleration in rectilinear motion. If the angular velocities at two instants of time at an interval T be ω_1 and ω_2 , then the mean rate of change of angular velocity in the interval is $\frac{\omega_2 - \omega_1}{T}$. If

now one considers the limit $T \rightarrow 0$, one gets the instantaneous rate of change of angular velocity, i.e., in other words, the angular acceleration.

As a special case, consider a circular motion with *uniform* angular acceleration. Here the instantaneous rate of change of angular velocity is the same as the average rate. As a consequence, if ω_1 and ω_2 be the angular velocities at two time instants at an interval of time, say, t , then

$$\omega_2 = \omega_1 + \alpha t, \quad (3-132)$$

where α stands for the angular acceleration.

The angle of rotation is now no longer given by (3-131) which, however, holds with ω replaced with the *average* angular velocity during the interval, this due to the fact that the angular acceleration is *uniform*:

$$\theta = \frac{\omega_1 + \omega_2}{2}t = \omega_1 t + \frac{1}{2}\alpha t^2, \quad (3-133)$$

where (3-132) has been made use of. Notice that equations (3-132) and (3-133) are, once again, analogous to corresponding equations for a rectilinear motion with uniform acceleration.

3.19.4 Centripetal acceleration

3.19.4.1 Uniform circular motion

Let us, to start with, consider a particle in uniform circular motion with a speed v as shown in fig. 3-23(A). Let r be the radius of the circle. Let, at time instants $t - \frac{\delta t}{2}$, t and $t + \frac{\delta t}{2}$, the particle be located at P, C and P' respectively, where δt denotes a small interval of time. Fig. 3-23(B) depicts the arc from P to P' in magnification, along with the radial lines PO, CO and P'O. The velocity vectors at the positions P and P' are shown with arrow-headed line segments, while CT denotes the tangent to the circle at C.

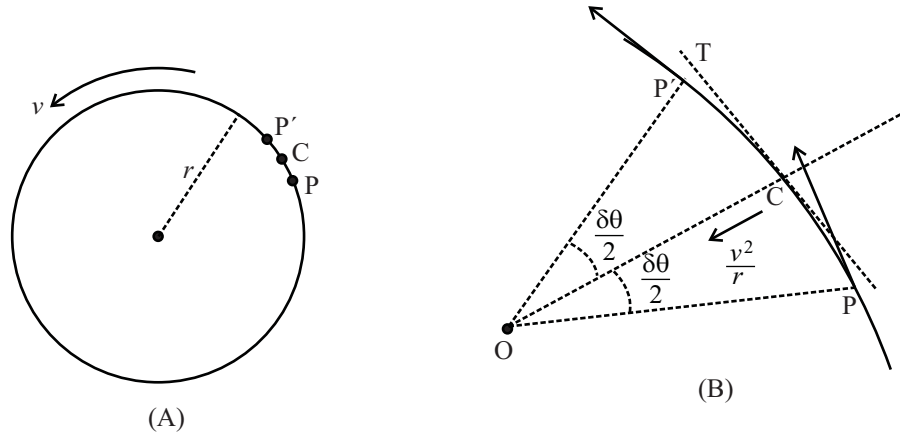


Figure 3-23: (A) The trajectory of particle in uniform circular motion; the radius of the circle is r ; P, C, P' are the positions of the particle at three successive time instants; (B) the arc PP', described in a small time interval, shown in magnification; CT is the tangent at C, while CO is the normal, O being the center of the circle; the velocities at P and P' are shown with arrows; the arc PCP' subtends an angle $\delta\theta$ at the center; the centripetal acceleration, of magnitude $\frac{v^2}{r}$, is directed from C to O.

Each of the velocity vectors at P' and P can be resolved into components along CT and CO respectively. In the present instance, the components along CT are equal, implying that there is no change in the velocity of the particle along the direction CT, the tangent to the point of location of the particle at time t , in the time interval δt . The components along CO, on the other hand, of the above two velocities are, respectively, $-v \sin(\frac{\delta\theta}{2})$ and $v \sin(\frac{\delta\theta}{2})$, where $\delta\theta$ stands for the angle subtended at the center of the circle by the arc PP' described in time δt . Since $\delta\theta (= \frac{v\delta t}{r})$, check this out) is a small angle for sufficiently small δt , the radial component of the change in the velocity of the particle in time δt is given by

$$\text{radial component of change in velocity} = \frac{v \sin \delta\theta}{r} = \frac{v^2 \delta t}{r}, \quad (3-134)$$

where, in the last equality, we have made use of the relation $\sin(\delta\theta) = \delta\theta$, which holds in the limit of infinitesimally small values of $\delta\theta$.

Thus, the radial component of the *rate* of change of velocity, i.e., of the *acceleration* of

the particle at the position C is given by

$$a_n = \frac{v^2}{r}, \quad (3-135)$$

where the subscript 'n' is used to designate the radial component of the acceleration which, in the present case of circular motion, is directed along the normal (pointing toward the center) to the circular trajectory at the instantaneous position P of the particle.

In other words, a particle in uniform circular motion having a speed v possesses, at any given instant, an acceleration in the radial direction given by eq. (3-135). This radial acceleration is proportional to the square of the velocity of the particle and, at the same time, is inversely proportional to the radius of the circular path. This acceleration along the radial direction is termed the *centripetal acceleration* of the particle.

If the angular velocity of the particle about the center O of the circular path be ω , then an alternative expression for the centripetal acceleration is (refer to eq. (3-130))

$$a_n = \omega^2 r \quad (3-136)$$

Since Newton's second law tells us that an acceleration of a particle in any given direction in an inertial frame is the result of some force or other acting along that direction, one concludes that a particle can possess a circular motion in an inertial frame only if a force acts on it along the radial direction at every instant. While such a force is at times referred to as a 'centripetal force', a more appropriate designation would be a 'force causing the centripetal acceleration'.

Problem 3-31

A car takes 300 s to cover a circular racing track of radius 2.0 km with uniform speed. What is its centripetal acceleration ?

Answer to Problem 3-31

HINT: The speed of the car is $v = \frac{2\pi \times 2.0 \times 10^3}{300} \text{ m}\cdot\text{s}^{-1}$. Hence the centripetal acceleration is $\frac{v^2}{r}$, where

$r = 2.0 \times 10^3 \text{ m}$ (answer: $0.88 \text{ m}\cdot\text{s}^{-2}$ (approx)).

Problem 3-32

Consider a small spherical ball tied to one end of a string of length l , the other end of which is attached to a fixed point P (fig. 3-24). The string is held in a horizontal position with the ball at rest, when the latter is released. As the string swings with the ball moving along an arc of radius l , the swing is obstructed by a peg Q vertically below O at a distance d . Find the condition for the string to swing completely round the peg, with the ball following a circular trajectory. Assuming that the string does swing completely, find the velocity of the ball as it reaches the highest point in its trajectory.

Answer to Problem 3-32

SOLUTION: The velocity v_0 of the ball as it reaches the lowest point in the circular trajectory of radius l is obtained from the principle of conservation of energy from the relation $\frac{1}{2}mv_0^2 = mgl$ (m = mass of the ball) since its potential energy decreases by mgl from the initial point (refer to eq (3-62)). Subsequently, the string starts to swing with the ball following a circular trajectory of radius $l - d$. As the ball rises through a height h from its lowest point (see fig. 3-24), its velocity v is given by $\frac{1}{2}mv_0^2 = \frac{1}{2}mv^2 + mgh$, again according to the principle of conservation of energy. The tension in the string (you will find below a brief introduction to the concept of tension in a string) at this point is given by $\frac{v^2}{l-d} = T + \frac{mg(h-(l-d))}{l-d}$, since the net force along the length of the string towards Q must provide for the centripetal acceleration.

Thus, the tension becomes zero for $h = l - \frac{d}{3}$ (check this out). If this height is $2(l - d)$, then the string becomes slack before it attains the vertical position with the ball above Q, and the ball then describes a parabolic trajectory as a projectile. On the other hand, if $h > 2(l - d)$, i.e., $d > \frac{3l}{5}$, then the string remains taut as the ball reaches the point Q' vertically above Q in its circular trajectory of radius $l - d$, and subsequently swings completely around the peg. The fact that this condition implies $d > \frac{l}{2}$, ensures that the ball possesses a non-zero velocity as it reaches Q'. According to the principle of conservation of energy, the velocity v' at Q' is given by $\frac{1}{2}mv_0^2 = \frac{1}{2}mv'^2 + 2mg(l - d)$, i.e., $v' = \sqrt{2g(2d - l)}$.

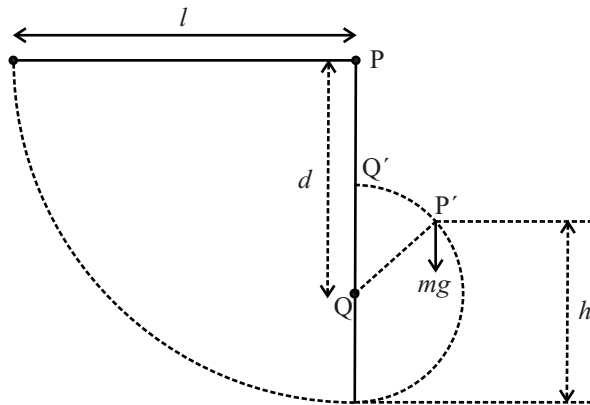


Figure 3-24: A small sphere attached to one end of a string of length l , the other end of which is attached to a fixed point P; if the ball is released from rest with the string held in a horizontal position, it follows a circular trajectory of radius l till the string gets caught in a peg at Q located vertically below P at a distance d ; the ball subsequently follows a circular trajectory of radius $l - d$. As the ball rises to a point P' through a height h from the lowest position, it attains a velocity v , the tension in the string being T ; if T remains positive as the ball reaches the point Q' at a height $l - d$ above Q, the string makes a complete swing round the peg with the ball following the circular trajectory; else the string becomes slack and the ball follows a parabolic trajectory as a projectile.

Digression: Tension in a string.

The tension in a string arises due to the force exerted by one segment of the string on a contiguous segment through the point separating the two. Thus, in fig. 3-25(A), the segment AB exerts a force T on the segment BC where the force acts at B in the direction shown by the arrow, while the segment BC exerts an equal and opposite force on AB (dotted arrow). One then says that the tension in the string at B is T . The tension in a string may be uniform, being the same at all points, or may vary from point to point. In the latter case, a segment of the string may experience a net force arising from the forces exerted on it by contiguous segments. Thus, in fig. 3-25(B), if the tension at the points P and Q in a string are T_1 and T_2 , which act along the tangents at P and Q respectively, then the segment PQ experiences a net force that is the resultant of the two forces of tension.

The forces responsible for tension are, in the ultimate analysis, of electromagnetic origin.



Figure 3-25: Illustrating the idea of tension in a string; (A) the segment AB of a string exerts a force T on the segment BC through the point B, while BC exerts an equal and opposite force (dotted arrow) on AB; the tension at B is then T ; (B) owing to the tension being non-uniform (i.e., varying from point to point on the string in magnitude or direction or both), the tension forces of magnitudes T_1 and T_2 acting on the segment PQ at P and Q may not balance, as a result of which a net force acts on the segment.

3.19.4.2 Motion along a space curve

The trajectory of a particle in three dimensions is, in general, a *space curve*. While space curves can be of various different descriptions, a *local* description of a space curve, i.e., one characterizing a small part of such a curve, involves only a few relevant parameters.

Consider, for instance, any point P on a space curve, with two other points (say, P_1 and P_2) on the curve lying on either side of P, in close proximity to it. As P_1 and P_2 are imagined to approach the point P to within an infinitesimally small distance from it, one obtains an infinitesimally small segment or arc of the space curve around P. In the process, the plane containing the three points P, P_1 , and P_2 approaches a limiting disposition, and in this limit, the plane is referred to as the *osculating plane* of the curve at P. To a good approximation, the curve can be assumed to be locally confined to this osculating plane.

Confining our attention to the small segment P_1PP_2 contained in the osculating plane, the tangent (PT, see fig. 3-26) to the curve drawn at P lies in this plane, while the normals P_1O , P_2O , again lying in this plane, intersect at the point O. Imagining the limiting situation in which P_1 and P_2 approach the point P, the limiting position of the point O is referred to as the *center of curvature* of the space curve with reference to the point P and the distance OP is termed the *radius of curvature* at P, the straight line PO being the *principal normal* (or, in brief, simply the *normal*) at P.

The local geometrical description of a space curve can be given in terms of the tangent

and the normal and, in addition, the *binormal* at the point P, which is a line perpendicular to the tangent and the principal normal. This local description involves a *rotation* about the binormal and a *torsion* about the tangent. However, the rate of torsion of the curve at P, though relevant in a local description of the curve, will not be of direct concern to us in the present context.

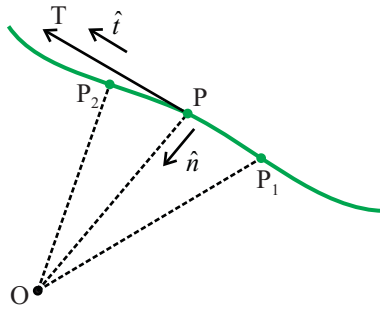


Figure 3-26: Illustrating centripetal acceleration for motion along a space curve; P_1 , P , and P_2 are three successive positions of the particle; \hat{t} denotes the unit tangent vector, directed along the tangent PT at P, while \hat{n} denotes the unit normal vector, directed towards the center of curvature O corresponding to the position P on the trajectory; the binormal and the unit vector (\hat{b}) along it are not shown; the centripetal acceleration at P is directed along PO; the plane of the figure is the osculating plane at P.

Let, at any given instant of time t , a particle be located at P on the space curve under consideration, the latter being its trajectory in a given motion, and let its speed at this instant be v where v may change with time. The velocity vector \mathbf{v} can then be written as $v\hat{t}$, where \hat{t} is a unit vector along the tangent PT to the curve, pointing in the direction of motion of the particle. Let the acceleration of the particle at time t be \mathbf{a} , where, in general, both \mathbf{v} and \mathbf{a} are time dependent vectors.

Denoting by \hat{n} and \hat{b} the unit vectors along the normal PO and the binormal (say, PB, not shown in fig. 3-26) respectively, one can resolve the acceleration \mathbf{a} of the particle at P into three components, namely, along \hat{t} , \hat{n} , and \hat{b} . Of these, the component of acceleration along \hat{n} is given by

$$a_n = \frac{v^2}{r}, \quad (3-137)$$

where r stands for the radius of curvature at P. If ω denotes the instantaneous angular velocity of the particle about the center of curvature O at time t , then one has the alternative expression

$$a_n = \omega^2 r. \quad (3-138)$$

Comparing with equations (3-135) and (3-136), a_n is termed the *centripetal acceleration* in this more general context.

In contrast to uniform circular motion, the speed, angular velocity and radius of curvature may, in general, all change with time. Also, in the general case, the centripetal acceleration is only one component of the instantaneous acceleration (the one along the normal PO) while in uniform circular motion a_n gives the magnitude of the total instantaneous acceleration of the particle, the entire acceleration being directed along the normal.

3.19.5 Radial and cross-radial accelerations in planar motion

While in sec. 3.19.4.2 I wrote down the expression for the centripetal acceleration in the case of a general motion of a particle along a space curve, a more restricted motion corresponds to a trajectory confined in a single plane.

For such a planar motion, which is of considerable interest in mechanics, the direction of the binormal remains unchanged with time and the acceleration at any given instant possesses only two components, namely, the one along \hat{t} (the unit vector along the tangent), and that along \hat{n} (the unit vector along the normal), where, in general, both \hat{t} and \hat{n} keep on changing with time. At the same time, the velocity vector at any given instant is directed along \hat{t} , with no component along \hat{n} .

One can also express the velocity and acceleration vectors in a planar motion in terms of another set of unit vectors, to be denoted below by \hat{e}_r and \hat{e}_θ , such a decomposition being at times a convenient one in the description of the planar motion.

In addition, the velocity (\mathbf{v}) and acceleration (\mathbf{a}) may be expressed in terms of unit vectors \hat{i} and \hat{j} along the x- and y-axes of any Cartesian co-ordinate system chosen in the plane of motion (these being special cases of equations (3-7) and (3-17) where the components of \mathbf{v} and \mathbf{a} along \hat{k} are both zero).

For this, we describe the planar motion in terms of *plane polar co-ordinates*, by choosing any point O in the plane as the origin, and any fixed line passing through O (say, OX in fig. 3-27) as a reference line.

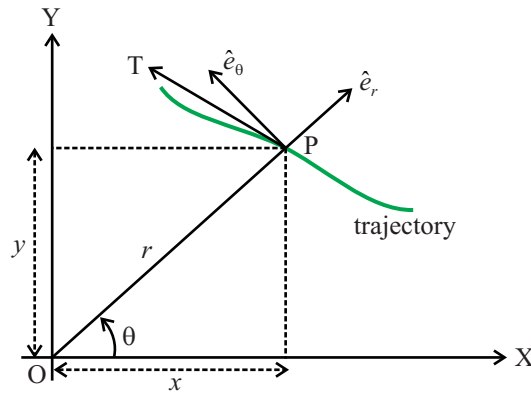


Figure 3-27: Illustrating plane polar co-ordinates with reference to the planar motion of a particle; P is the instantaneous position of the particle on the planar trajectory; the unit vector \hat{e}_r is directed along OP, where O is the origin of a rectangular co-ordinate system with OX and OY as axes; the plane polar co-ordinates of P are (r, θ) , where $OP = r$, and θ is measured counter-clockwise from OX, the latter being chosen as the reference line; the unit vector \hat{e}_θ is orthogonal to \hat{e}_r and points in the direction of increasing θ ; the polar co-ordinates (r, θ) are related to the Cartesian co-ordinates (x, y) as in eq. (3-141).

For any point P in the plane, its plane polar co-ordinates r and θ are defined as in the figure, where r is the distance from the origin O and θ is the angle $\angle POX$ (measured in the anti-clockwise sense from OX to OP as shown in the figure by the curved arrow). One then defines \hat{e}_r as the unit vector along OP, and \hat{e}_θ as the unit vector perpendicular to \hat{e}_r in the direction of θ increasing (fig. 3-27).

Considering now the planar motion of a particle referred to the plane polar co-ordinates defined above, if (r, θ) denote the plane polar co-ordinates of the particle at any given time t , then the instantaneous position vector with respect to the origin O (which is

directed along OP and has a magnitude r) is given by

$$\mathbf{r} = r\hat{e}_r. \quad (3-139)$$

The instantaneous velocity and acceleration vectors of the particle, on the other hand, have, in general, components both along \hat{e}_r and \hat{e}_θ , these being referred to as the *radial* and *cross-radial* components of these vectors respectively. The relevant expressions for \mathbf{v} and \mathbf{a} are

$$\mathbf{v} = \dot{r}\hat{e}_r + r\dot{\theta}\hat{e}_\theta, \quad \mathbf{a} = (\ddot{r} - r\dot{\theta}^2)\hat{e}_r + (2\dot{r}\dot{\theta} + r\ddot{\theta})\hat{e}_\theta, \quad (3-140)$$

where a dot over the symbol representing a time dependent quantity is used to denote the instantaneous time derivative of that quantity, while a double dot denotes the second order time derivative.

In this context, it is worthwhile to have a look at the relation between polar and Cartesian co-ordinates, where both the co-ordinate systems are chosen to have a common origin (O) and the reference line of the polar system is chosen along the x-axis of the Cartesian system, as in fig. 3-27. For any point P with polar co-ordinates r, θ and Cartesian co-ordinates x, y , one then has

$$x = r \cos \theta, \quad y = r \sin \theta. \quad (3-141)$$

At the same time, the unit vectors $\hat{e}_r, \hat{e}_\theta$ are related to \hat{i}, \hat{j} as

$$\hat{e}_r = \hat{i} \cos \theta + \hat{j} \sin \theta, \quad \hat{e}_\theta = -\hat{i} \sin \theta + \hat{j} \cos \theta. \quad (3-142)$$

Notice that, while the unit vectors \hat{i}, \hat{j} do not depend on the position of the point P in the plane, \hat{e}_r and \hat{e}_θ do depend on the position of P. In this, the unit vectors \hat{e}_r and \hat{e}_θ resemble the unit tangent vector \hat{t} and unit normal vector \hat{n} for a plane curve, where the latter, moreover, are defined only in relation to the curve.

It is straightforward to see that the expression for \mathbf{a} in eq. (3-140) implies the expres-

sion (3-135) for the centripetal acceleration of a particle in uniform circular motion.

3.19.6 Angular momentum

3.19.6.1 Angular momentum about a point

Consider a particle located at any given instant of time at P, whose position vector with respect to a chosen origin O is, say, \mathbf{r} . If \mathbf{p} denotes the instantaneous momentum of the particle, then the *angular momentum* (also referred to as the *moment of momentum*) of the particle about O at the instant under consideration is defined as

$$\mathbf{L} = \mathbf{r} \times \mathbf{p} = m\mathbf{r} \times \mathbf{v}, \quad (3-143)$$

where m denotes the mass and \mathbf{v} the instantaneous velocity of the particle. The angular momentum is a vector whose direction is perpendicular to both \mathbf{r} and \mathbf{p} , being related to the directions of these two by the right hand rule. Its magnitude is given by

$$L = mrv \sin \theta = m\rho v, \quad (3-144)$$

where θ denotes the angle between \mathbf{r} and \mathbf{v} , and ρ stands for the length of the perpendicular dropped from O on the line passing through P along the direction of motion of the particle (check this out).

The unit of angular momentum in the SI system is $\text{kg}\cdot\text{m}^2\cdot\text{s}^{-1}$.

From equations (3-126) and (3-143) one finds that the angular momentum is related to the instantaneous angular velocity about O as

$$\mathbf{L} = mr^2\vec{\omega}. \quad (3-145)$$

In mechanics, the use of angular momentum is often of greater convenience than that of angular velocity, just as the use of momentum (sometimes referred to as *linear momentum*) offers a greater advantage compared to that of velocity.

3.19.6.2 Angular momentum about an axis

Analogous to the definition of angular velocity about an axis, one can also define the angular momentum of a particle about an axis. Considering, for instance, fig. 3-20, one defines the instantaneous angular momentum about the axis AB as

$$\vec{\Lambda} = ((\vec{\rho} \times \mathbf{p}) \cdot \hat{n})\hat{n}, \quad (3-146a)$$

or, equivalently, as

$$\vec{\Lambda} = m\rho^2\vec{\Omega}, \quad (3-146b)$$

where $\vec{\Omega}$ stands for the angular velocity about the axis under consideration, and ρ is the perpendicular distance of the instantaneous position of the particle from the axis.

The angular momentum about AB is related to the angular momentum about any point O chosen to lie on AB as

$$\vec{\Lambda} = (\mathbf{L} \cdot \hat{n})\hat{n}, \quad (3-146c)$$

i.e., the angular momentum about the axis AB is the *projection* along AB of the angular momentum about O.

1. While \mathbf{L} depends on the choice of O on the axis, its projection does not.
2. Recall that there can be two unit vectors \hat{n} parallel to any given line AB, directed oppositely to each other. In equations (3-146a) and (3-146c), \hat{n} can be taken to be any one of these or, specifically, as the one inclined at an acute angle to $\vec{\rho} \times \mathbf{p}$. In the former case, the scalar component of $\hat{\Lambda}$ along \hat{n} can be either positive or negative while, in the latter, it will be a positive quantity.

Evidently, the magnitude (Λ) of the angular momentum about AB is related to that about the point O as

$$\Lambda = L \cos \phi, \quad (3-146d)$$

where ϕ is the angle shown in fig. 3-21. This equation contrasts with the relation (eq. (3-129a) or, in the special case of a rotational motion about the axis AB, eq. (3-129b)), between the angular *velocity* about AB and that about O.

Problem 3-33

Consider an axis AB passing through the origin O of a Cartesian co-ordinate system and making an acute angle $\cos^{-1} \frac{1}{\sqrt{3}}$ with each of the axes, and a point particle of mass 0.2 kg located at the point P with co-ordinates (1, 0, 0). If the instantaneous velocity of the particle is (in $\text{m}\cdot\text{s}^{-1}$) $\mathbf{v} = 2\hat{i} - \hat{j}$, find the angular velocity of the particle about O and about AB. Verify the relation (3-129a). Obtain the angular momenta about O and AB as well, and verify eq. (3-146d).

Answer to Problem 3-33

HINT: Choose $\hat{n} = \frac{1}{\sqrt{3}}(\hat{i} + \hat{j} + \hat{k})$ as one of the two possible unit vectors parallel to the given axis. Here (SI units are implied for all the physical quantities) $\vec{\omega} = \frac{\mathbf{r} \times \mathbf{v}}{r^2} = -\hat{k}$. Further, $\vec{\rho} = \mathbf{r} - (\mathbf{r} \cdot \hat{n})\hat{n} = \frac{2}{3}\hat{i} - \frac{1}{3}\hat{j} - \frac{1}{3}\hat{k}$, and $\vec{\rho} \times \mathbf{v} = -\frac{1}{3}\hat{i} - \frac{2}{3}\hat{j}$. Thus, $\vec{\Omega} = \frac{(\vec{\rho} \times \mathbf{v}) \cdot \hat{n}}{\rho^2} \hat{n} = -\frac{1}{2}(\hat{i} + \hat{j} + \hat{k})$, $\Omega = \frac{\sqrt{3}}{2}$. Since $\cos \phi$ is the cosine of the angle between $\vec{\omega}$ and $\vec{\Omega}$, we have $\omega \cos \phi = \frac{1}{\sqrt{3}}$. On the other hand, β being the angle between \mathbf{r} and either of the two directed lines parallel to AB, $\sin \beta = \sqrt{\frac{2}{3}}$, and $\Omega \sin^2 \beta = \frac{1}{\sqrt{3}}$. This verifies eq. (3-129a).

The angular momentum about O is (SI units implied) $\mathbf{L} = m\mathbf{r} \times \mathbf{v} = -0.2\hat{k}$, and $L \cos \phi = \frac{0.2}{\sqrt{3}}$. Again $\Lambda = m\rho^2\Omega = 0.2 \times \frac{2}{3} \frac{\sqrt{3}}{2} = \frac{0.2}{\sqrt{3}}$, which verifies eq. (3-146d).

3.19.6.3 Angular momentum in circular motion

In the special case of a circular motion, the magnitude of the angular velocity about the center of the circle is related to the speed along the circumference by eq. (3-130), the direction of the angular velocity being along the perpendicular to the plane of the circle, related to the sense of rotation by the right hand rule. Analogously, the magnitude of the angular momentum about the center of the circle is given by

$$L = mav = ma^2\omega, \quad (3-147)$$

where a stands for the radius of the circle. The direction of the angular momentum is the same as that of the angular velocity (see eq. (3-145)).

The magnitude and direction of the angular momentum about an axis passing through the center of the circle and perpendicular to its plane are the same as those about the origin since, in this case, \mathbf{L} is parallel to the axis (see eq. (3-146c)).

3.19.7 Moment of a force

3.19.7.1 Moment of a force about a point

In fig. 3-28, O is any given origin, and AB is the line of action of a force \mathbf{F} acting on a particle or a rigid body. Let P be any point on the line of action of the force. One can take it as the point of application of the force.

1. For a force acting on a particle, while it is customary to take the particle itself as the point of application of the force, the latter can more generally be taken as any point lying on the line of action of the force and imagined to be rigidly connected to the particle. A similar statement applies to a force acting on a rigid body.
2. It can be seen from the following definition of the moment of the force \mathbf{F} about the point O that the choice of the point P on the line of action of the force does not really matter.

Denoting the position vector of P with respect to O as \mathbf{r} , the *moment of the force \mathbf{F}* about O is defined by the expression

$$\mathbf{M} = \mathbf{r} \times \mathbf{F}. \quad (3-148a)$$

The moment is a vector directed along the normal (the line OM in fig. 3-28) to the plane containing the point O and the line of action AB , and is related to the directions of \mathbf{r} and \mathbf{F} by the right hand rule. If N be the foot of the perpendicular dropped from O on AB , and the vector extending from O to N is denoted by $\vec{\rho}$, then an equivalent expression for

M is

$$\mathbf{M} = \vec{\rho} \times \mathbf{F}, \quad (3-148b)$$

which tells us that the moment does not actually depend on the position of the point P on AB , i.e., what is relevant here is the line of action of the force. In particular, the magnitude of M is given by

$$M = \rho F. \quad (3-148c)$$

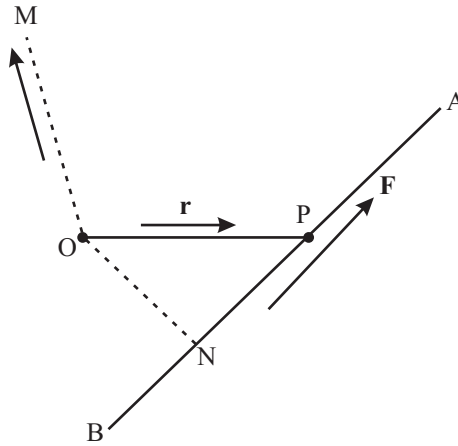


Figure 3-28: Illustrating the concept of moment of a force about a point O ; the force \mathbf{F} on a particle at P acts along the line AB ; N is the foot of the perpendicular dropped from O on AB ; denoting by $\vec{\rho}$ the vector extending from O to N , the moment is defined as $\vec{\rho} \times \mathbf{F}$, or, equivalently, as $\mathbf{r} \times \mathbf{F}$, though the moment does not depend on the position of P on AB .

For a given line of action of the force, the torque depends on the choice of the origin O . The moment is *zero* if O lies on the line of action of the force.

Another name for the moment of a force is *torque*. The unit of torque in the SI system is N·m. Though formally the same as the joule, it is not usually written as J so as to distinguish it from work or energy.

Combining the definitions of the angular momentum of a particle about a point and of the moment of the force acting on the particle about the same point, one arrives at an

important conclusion: *the rate of change of angular momentum is the torque*, i.e.,

$$\frac{d\mathbf{L}}{dt} = \mathbf{M}. \quad (3-149)$$

This shows that *the effect of torque is to change the state of angular motion* by changing the angular momentum. This is analogous to the role of force, which changes the state of motion by changing the linear momentum of a system.

At times, a torque is described with a *signed number* (with the appropriate unit) instead of using a vector. Considering the plane containing the point O and the line of action of the force, one chooses any one side of this plane as the reference side. For instance, among the two sides of the plane A in fig. 3-29, we choose the one above A as the reference side. If the direction (or, more precisely, the *sense*) of the force, with reference to the point O be *anticlockwise* as looked at from the reference side, then the torque is described with a *positive* sign attached to its magnitude. If, on the other hand, the sense of the force with reference to O be *clockwise* as looked at from the reference side, then a *negative* sign is attached to its magnitude. For instance, the torque of the force F is positive in fig. 3-29(A), and negative in 3-29(B).

While a vector quantity like velocity, momentum, or force has, associated with it, a *direction*, a quantity like angular velocity, angular momentum, or torque has, associated with it, a *sense of rotation*. Such quantities go by the name of pseudo-vectors or *axial* vectors, in contrast to vectors like velocity and momentum which are distinguished from axial vectors by referring to these as *polar* vectors. An axial vector like the angular velocity or torque is expressed in the form of a vector product of two polar vectors. The sense of rotation from the first of these two polar vectors to the second is the sense associated with the vector product. The direction related to this sense of rotation by the right hand rule is then chosen as the direction of the axial vector (refer to sec. 2.6.1).

3.19.7.2 Moment of a force about an axis

Similar to the definition of the angular momentum or *moment* of momentum about an axis, one can also define the moment of a force about any given axis as

$$\vec{\Gamma} = ((\vec{\rho} \times \mathbf{F}) \cdot \hat{n})\hat{n}, \quad (3-150)$$

where \hat{n} is a unit vector along the axis, while $\vec{\rho}$ is also defined as before: considering any point P on the line of action on the force and dropping a perpendicular PN from P to the axis, $\vec{\rho}$ is the vector extending from N to P. Choosing any point O on the axis, the moment about the axis ($\vec{\Gamma}$) is seen to be simply the projection on the axis of the moment (\mathbf{M}) about O:

$$\vec{\Gamma} = (\mathbf{M} \cdot \hat{n})\hat{n}. \quad (3-151)$$

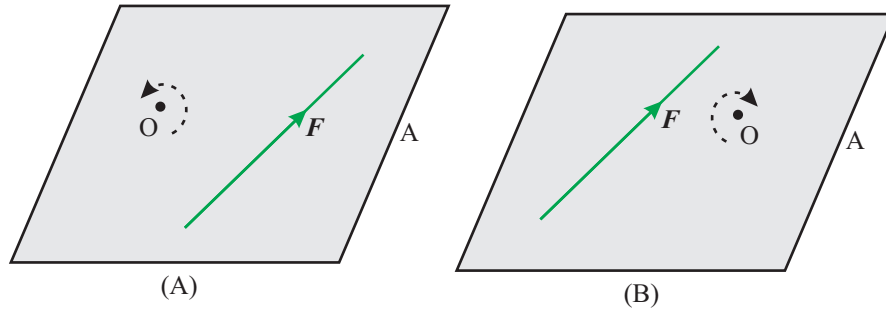


Figure 3-29: Sign of torque; looked at from above the plane A containing the point O and the line of action of the force \mathbf{F} , the torque about O is positive (anticlockwise) in (A), and negative (clockwise) in (B).

Making use of eq. (3-149) and of the relations between the angular momentum and torque about a point on the one hand and the corresponding quantities about an axis on the other, one finds that, given an axis AB, the instantaneous rate of change of angular momentum of a particle about any given axis is equal to the moment of the

forces acting on the particle about that axis:

$$\frac{d\vec{\Lambda}}{dt} = \vec{\Gamma}. \quad (3-152)$$

3.19.7.3 Impulse of a torque

Analogous to the impulse of a force acting on a particle, one can define the impulse of a torque acting on a particle in its angular motion about a given point in terms of the expression

$$\mathbf{I} = \int \mathbf{M} dt, \quad (3-153)$$

where the integration is over any given interval of time. One can then relate the impulse to the change of angular momentum about the point under consideration during that interval,

$$\mathbf{I} = \mathbf{L}_2 - \mathbf{L}_1. \quad (3-154)$$

Here \mathbf{L}_1 , \mathbf{L}_2 stand for the angular momentum about the given point at the beginning and end of the time interval under consideration.

It is straightforward to write down the corresponding results for angular motion about any given axis:

$$\mathbf{I} = \int \vec{\Gamma} dt, \quad (3-155)$$

$$\mathbf{I} = \vec{\Lambda}_2 - \vec{\Lambda}_1, \quad (3-156)$$

where the symbols are self-explanatory.

In numerous situations of practical interest, it makes sense to talk of an *impulsive torque*, where the torque acts for a short time interval on the particle under considera-

tion, still producing a non-zero impulse. For such an impulsive torque, the impulse of the torque is more convenient to work with than the torque itself.

3.19.8 Angular motion of a system of particles

Considerations relating to the angular motion of a particle can be extended to those involving a system of, say, N particles. In particular, the total angular momentum of the system of particles with instantaneous position vectors $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N$ relative to a chosen origin O , and with instantaneous momenta $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_N$ respectively, is

$$\mathbf{L} = \mathbf{r}_1 \times \mathbf{p}_1 + \mathbf{r}_2 \times \mathbf{p}_2 + \dots + \mathbf{r}_N \times \mathbf{p}_N = \sum_{i=1}^N \mathbf{r}_i \times \mathbf{p}_i, \quad (3-157)$$

about the point O , where the summation symbol has been used for the sake of brevity.

Analogous to the definition of the angular momentum of a particle about any given axis, say, AB , one can define the angular momentum of a system of particles about the axis as well. One first works out the angular momentum of each particle about any chosen origin, say, O on AB , and then sums up the projections of all these along AB . Equivalently, one can sum up expressions similar to eq. (3-146a).

The basic equation describing the angular motion of a system of particles follows from eq. (3-149) by summing up over all the particles the system is made of,

$$\frac{d\mathbf{L}}{dt} = \mathbf{M}, \quad (3-158)$$

which looks similar to eq. (3-149), but with an altered meaning of the symbols: \mathbf{L} and \mathbf{M} now stand for the total angular momentum and total moment of the forces about a chosen origin.

In calculating the total moment of the forces acting on the particles of a given system of particles, one need not take into consideration the internal forces if these internal forces are central in nature, because the moments of the internal forces cancel pairwise by virtue of Newton's third law.

One can also write down, for the given system of particles and the forces acting on these, the equation, analogous to eq. (3-152), relating the total moment ($\vec{\Gamma}$) of the forces about any given axis, and the rate of change of the total angular momentum ($\vec{\Lambda}$) about that axis,

$$\frac{d\vec{\Lambda}}{dt} = \vec{\Gamma}, \quad (3-159)$$

where, once again, only the external forces acting on the system under consideration are to be considered in working out the total moment Γ .

One can express eq. (3-158) (resp. eq. (3-159)) in terms of the total impulse of the external torques about the chosen origin (resp. about the given axis) acting on the system under consideration. The result looks like eq. (3-154) (resp. eq (3-156)) with an altered significance of the symbols which now correspond to a system of particles rather than to a single particle.

3.19.9 Principle of conservation of angular momentum

Equation (3-158) which can be looked upon as the basic equation describing the angular motion of a system of particles about any given point O, implies the *principle of conservation of angular momentum*: if the total moment about O of the external forces acting on the system be zero, the total angular momentum about O is conserved. What is notable here, is that the forces *internal* to the system are not to be considered, since their total moment vanishes by virtue of Newton's third law (assuming that these internal forces are central in nature).

An alternative form of the principle of conservation of angular momentum is obtained by referring to the angular motion about any axis AB where, once again, the angular momentum of the system under consideration about the axis is conserved if the total moment of the external forces about the axis vanishes.

The principle can also be stated in terms of the total impulse of the external torques

(about a given point or a given axis, as the case may be) on the system under consideration: if the impulse over a given interval of time vanishes then the total angular momentum of the system (about the point or the axis) at the beginning of the interval will be the same as that at the end.

The principle of conservation of angular momentum is of especial importance in describing the *rotational* motion (see sec. 3.19.10) of rigid bodies about given axes.

Problem 3-34

Consider two particles A and B of masses m_1 and m_2 respectively, attached to a light and rigid rod of length l at the two ends of the latter, which is capable of rotation about a horizontal axis perpendicular to its length passing through its center O (see fig. 3-30). With the rod stationary in a horizontal position, a particle C of mass m descends vertically with a velocity v on top of A and gets attached to it. What is the angular velocity imparted to the rod?

Answer to Problem 3-34

HINT: Considering the closed system made up of the rod and the three particles, the principle of conservation of angular momentum applies in that the total angular momentum of the system about the axis just before the impact of C on A has to be the same as that immediately after the impact. The external forces acting on the system here are the forces of gravity on the particles, but the total impulse of the torques of these forces, considered for the vanishingly small interval of the impact, is zero (the forces are finite while the relevant time interval goes to zero). One thus has $mv\frac{l}{2} = (m_1 + m)\omega(\frac{l}{2})^2 + m_2\omega(\frac{l}{2})^2$, where ω is the required angular velocity just after the impact. This gives $\omega = \frac{2mv}{(m_1 + m_2 + m)l}$.

Linear and angular motions: manifestations of the same basic phenomenon.

When one speaks of the linear and angular motions of a particle or of a system of particles, one does not actually refer to two distinct motions because the two are, in reality, descriptions of the same basic motion of the system under consideration.

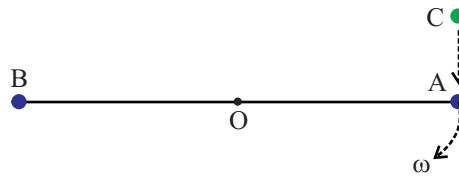


Figure 3-30: A light rod, capable of rotation about a horizontal axis passing through its center O (and perpendicular to its length), with point masses A and B attached to its ends; with the rod resting in a horizontal position, a particle C descends vertically on A and sticks to it; the rod thereby acquires an angular momentum ω about O.

For instance, look at fig. 3-9 showing a part of the trajectory of a particle, on which A denotes the instantaneous position of the particle and AB denotes the tangent to the trajectory at the point A, where the instantaneous velocity and momentum of the particle are directed along AB. One then says that AB gives the instantaneous direction of *linear motion* of the particle.

Consider now any other point O or axis CD (not shown in the figure). The particle, in the course of its motion, describes a certain angle, say, $\delta\phi$, about the point O or about the axis CD in any small interval of time δt . This one describes as the angular motion (with reference to O or to CD, as the case may be) of the particle in the interval δt . Thus, the term angular motion is necessarily associated with a reference point or an axis. Instead of the point O or the axis CD, one could describe the same motion with reference to some other point or some other axis, in which case the *rate* at which the particle describes the angle would differ. The basic phenomenon, however, remains the same, namely the motion of the particle along the trajectory under consideration.

3.19.10 Rotational motion about an axis: moment of inertia

Of especial importance is the case of instantaneous *rotational motion* about the axis AB. Recall that the instantaneous motion of a particle is said to be one of rotation about a given axis if its instantaneous velocity is perpendicular to the plane containing the axis and the instantaneous position of the particle. The angular velocity of rotation (ω) in this case is the angular velocity of the particle about the foot of the perpendicular dropped from the instantaneous position of the particle on the axis. If the distance of the particle from the foot of the perpendicular be a , then the instantaneous velocity of the particle

is given by (see eq. (3-127))

$$v = a\omega. \quad (3-160)$$

If, now, the instantaneous motion of *all* the particles of a system be one of rotation about an axis AB, and if the angular velocities of rotation of all of these about AB be the same, say $\vec{\omega}$ (note the change of notation compared to above, where the angular velocity about an axis has been denoted by $\vec{\Omega}$; the angular momentum about the axis will be similarly denoted by \vec{L} instead of $\vec{\Lambda}$), then the instantaneous motion of the system of particles will be said to be one of rotation about AB. The magnitude of angular momentum about the axis is then given by the expression

$$L = \left(\sum_{i=1}^N m_i \rho_i^2 \right) \omega, \quad (3-161a)$$

where m_i stands for the mass of the i th particle ($i = 1, 2, \dots, N$), and ρ_i for the perpendicular distance of the i th particle from the axis (check this equation out). In this expression, ω and L can be looked upon as scalar components of the corresponding vectors along any one of the two unit vectors parallel to AB, or else, they can equivalently be considered as quantities with appropriate signs related to the sense of rotation. As I have indicated above, looking along the axis in any of the two possible directions as the reference direction, ω and L will be taken to be positive for an anticlockwise rotation, and negative for a clockwise rotation (with reference to the chosen direction).

The quantity $\sum m_i \rho_i^2$ appearing in the expression (3-161a) is referred to as the *moment of inertia* of the system of particles under consideration about the chosen axis AB (see fig. 3-31 for illustration). Recall that the motion of the system of particles has been assumed to be one of rotation at any particular instant of time. At a later instant, the motion may not be found to be one of rotation, or may be a rotation with different values of the distances ρ_i and of ω . Thus, in general, the moment of inertia of a system of particles about any chosen axis need not be a constant.

Denoting the moment of inertia of the system of particles about the axis AB by the

symbol I ,

$$I = \sum_{i=1}^N m_i \rho_i^2, \quad (3-161b)$$

the above expression for angular velocity assumes the form

$$L = I\omega. \quad (3-161c)$$

This is analogous to the relation between the linear momentum, mass, and velocity of a particle or, more generally, to the relation between the corresponding quantities in the center of mass motion of a system of particles.

Note that eq. (3-161c) makes a statement only about the component of the angular momentum about the axis of rotation (where a rotational motion is assumed) and not about the angular momentum vector (\mathbf{L}) about any given point on the axis which, in general, points in a direction other than the angular velocity, the latter being directed along the axis of rotation. If, however, the axis of rotation happens to be an *axis of symmetry* of the system of particles or body under consideration, then the vector \mathbf{L} also points along the axis.

Incidentally, eq. (3-159), which is the central formula for describing the angular motion (including rotational motion) of a system of particles about an axis, is sometimes written with a changed notation, with \mathbf{L} (instead of \vec{L}) denoting the angular momentum about the axis (the same symbol \mathbf{L} is also used to denote the angular momentum about a point, depending on the context). One may thus find the equation written in the form, say,

$$\frac{d\mathbf{L}}{dt} = \mathbf{N}, \quad (3-162)$$

where \mathbf{N} (instead of $\vec{\Gamma}$) now denotes the total moment of the external forces about the axis under consideration (see, e.g., equations (3-163) and (3-164) below).

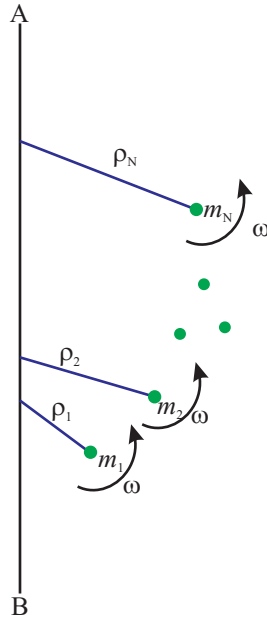


Figure 3-31: Illustrating the concept of moment of inertia of a system of particles about an axis; if all the particles happen to execute a rotational motion about the axis with a common angular velocity ω , then the angular momentum assumes the simple form $I\omega$, where I is the moment of inertia.

In a rotational motion about the axis under consideration, the instantaneous rates of change of the distances ρ_i ($i = 1, 2, \dots, N$) of the particles from the axis are all zero, and hence, from eq. (3-161b), the instantaneous rate of change of the moment of inertia I is also zero. One then has

$$I \frac{d\omega}{dt} = N, \quad (3-163)$$

where N and ω denote the scalar components of the corresponding vectors along the axis (each being represented by a magnitude along with a sign).

More generally, the moment of inertia of a system about the axis under consideration may change with time while each of the constituent particles shares a common angular velocity about the axis, in which case one has to consider the equation

$$\frac{d(I\omega)}{dt} = N. \quad (3-164)$$

If the moment N of the external forces about the axis vanishes, then the principle of conservation of angular momentum assumes the form

$$I\omega = \text{constant.} \quad (3-165)$$

Imagine, for instance, a child standing on a freely rotating platform (with a vertical axis of rotation) with her arms stretched horizontally on the two sides (see fig. 3-32(A)). If, now, she lowers her arms down to a vertical position, then the moment of inertia of the system about the axis of rotation will decrease (reason this out) and hence the angular velocity of the platform will increase so as to keep the product $I\omega$ constant. Note, however, that the kinetic energy of rotation $\frac{1}{2}I\omega^2 = \frac{(I\omega)^2}{2I}$ (see eq. (3-168)) will *increase* in the process.

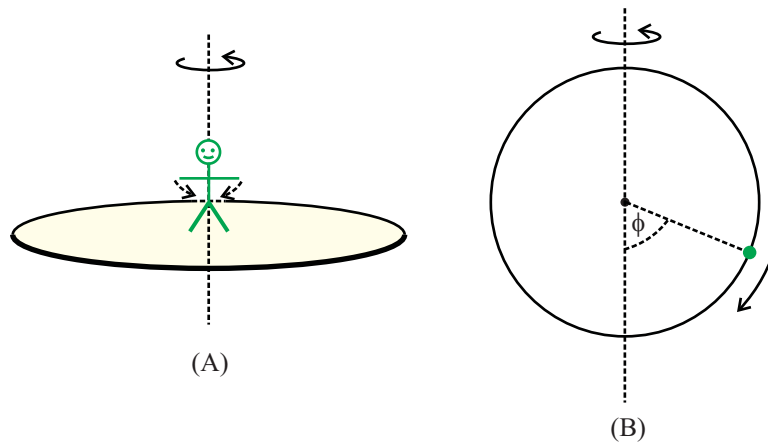


Figure 3-32: A system capable of executing rotational motion about an axis, where the moment of inertia can change with time; (A) a child on a freely rotating platform, lowering her stretched arms from an extended horizontal position down to a vertical position; (B) a weightless hoop rotating freely about a fixed vertical axis, on which a massive bead can slide without friction, causing the angle ϕ to decrease (ϕ may increase during part of the motion as well); in either case, the angular velocity of rotation about the vertical axis increases, as required by the principle of conservation of angular momentum and, at the same time, the kinetic energy associated with the rotational motion increases; the energy balance equation explaining this increase in kinetic energy may involve various factors, depending on the system under consideration.

This increase in the kinetic energy may be caused by various factors, depending on the system under consideration. For instance, consider a weightless circular hoop capable of rotating freely about a fixed vertical axis (with the plane of the hoop being

vertical at each instant, see fig. 3-32(B)), there being a bead (a point mass, for the sake of simplicity) mounted on it that can slide freely along its circumference. If the bead is released from a position such that the angle ϕ shown in fig 3-32(B) is, say, $\frac{\pi}{2}$, then it will descend down the circumference of the hoop, causing the moment of inertia to decrease and the angular velocity of rotation of the hoop to increase. In explaining the energy balance of the system in this case, one will have to consider, in addition to the kinetic energy of rotation (K_1) of the hoop and the bead about the vertical axis, and the gravitational potential energy of the bead, the kinetic energy (K_2) associated with the motion of the bead in the instantaneous plane of the hoop as well. On considering the total energy, it may be seen that the increase in the kinetic energy of rotation about the vertical axis is balanced by a corresponding decrease in the gravitational potential energy of the bead and its kinetic energy of instantaneous motion in the vertical plane.

Or, it may even be the case that the increase in rotational kinetic energy (about the vertical axis) is caused by some *other* source of energy. This additional source of energy may be an *internal* or an external one. In the example of the bead sliding on the hoop, the force of reaction exerted on the bead by the hoop is of the nature of an internal force which, however, *does not* perform any work and need not be considered in the energy balance. On the other hand, if one considers the frictional force on the bead, then the work done against the latter is to be considered in accounting for the energy balance. The frictional force in this instance is an internal one that does perform work (and constitutes an example of a *dissipative* force).

In the case of the child on the rotating platform, the increase in rotational kinetic energy is accounted for, in part, by the work done by her muscles as she continues to rotate along with the platform while, at the same time, lowering her arms or moving those in some other manner. The force exerted by her muscles is again an internal one that performs work during the motion.

In other words, while the principle of conservation of angular momentum is of general validity provided the moments of all forces about the axis of rotation add up to zero, the energy balance will involve work done by forces that may be of an internal and dissipative nature. This is analogous to the case of inelastic collisions considered in sec. 3.17.7.2 where the principle of conservation of momentum is of general validity

(in the absence of external forces) but the energy accounting has to take into consideration the transformation between kinetic energy and the energy of internal modes.

3.19.11 Work done in rotational motion

Suppose that the instantaneous motion of a system of particles is one of rotation about an axis AB, and consider a small interval of time during which the angle of rotation of the system about the axis is $\delta\phi$. Let the resultant of the moments, about the axis AB, of the forces acting on the particles at the instant under consideration be $N\hat{n}$, where \hat{n} is the unit vector along AB in a direction related to the sense of rotation of the particles by the right hand rule (refer to sec. 2.8). The work done by the system of forces in the infinitesimal rotation is then given by

$$\delta W = N\delta\phi. \quad (3-166)$$

Problem 3-35

Derive the relation (3-166).

Answer to Problem 3-35

If $\vec{\rho}_i$ denotes the vector distance of the i th particle of the system ($i = 1, 2, \dots, N$) from the axis AB, and \mathbf{F}_i the instantaneous force on the particle, then

$$\delta W = \sum_i \mathbf{F}_i \cdot \delta \mathbf{r}_i = \sum_i \mathbf{F}_i \cdot \delta\phi \hat{n} \times \vec{\rho}_i = \delta\phi \hat{n} \cdot \sum_i \vec{\rho}_i \times \mathbf{F}_i = N\delta\phi$$

.

Considering a Cartesian system of co-ordinates such that the instantaneous axis AB passes through the origin of this system, one can write

$$\delta W = \sum_{i=1}^3 N_i \delta\phi_i, \quad (3-167)$$

where N_i ($i = 1, 2, 3$) are the components of $\mathbf{N}(\equiv N\hat{n})$ along the three co-ordinate axes, and $\delta\phi_i$ are similarly the three components of the vector angle of rotation $\delta\phi\hat{n}$. N_i can be interpreted as the resultant instantaneous moment of the forces about the i th axis. The infinitesimal rotational motion with vector rotation angle $\delta\phi\hat{n}$ can be interpreted as the resultant of three infinitesimal rotations, with rotation angles $\delta\phi_i$ about the axes.

The expression (3-167) for a rotational motion is analogous to the expression for work done in the translational motion of a particle (or, more generally, of the center of mass motion of a system of particles), in terms of the components of the force acting on it and the components of its displacement with respect to any Cartesian co-ordinate system. In this sense, the moments N_i (note that the index i refers here to a Cartesian component and not to a particle in the system under consideration) are sometimes referred to as the *generalized forces* acting on the system of particles. The components of the vector angle of rotation and those of the angular momenta are similarly referred to as the changes in *generalized co-ordinates*, and the *generalized momenta* of the system.

Eq. (3-167) expresses the fact that, the general form of the expression for work involves the generalized forces and the changes in the generalized co-ordinates. This form of the expression for work holds for other, more general, systems as well and is of considerable significance in physics. For instance, for a broad class of systems, it serves to *define* the concept of *generalized force* in terms of that of work. Such a generalized force for a system cannot, in general, be said to act on this or that particle of the system. Moreover, it may not be of the dimension of newton, as in the case of a force that features in Newton's second law. In the case of angular motion, it has the dimension N-m, since it is some component of a torque. In other instances, it can even be of some other dimension.

Analogous to the result that the work done on a particle or a system of particles in translational motion equals the increase in its kinetic energy, one can state that the work done in rotational motion about an axis equals the increase in rotational kinetic

energy, where the rotational kinetic energy at any instant can be expressed in the form

$$K_{\text{rot}} = \frac{1}{2}I\omega^2. \quad (3-168)$$

In this expression, I stands for the moment of inertia of the particle or the system of particles about the axis under consideration, and ω denotes the instantaneous angular velocity.

Problem 3-36

Check the relation (3-168) out.

Answer to Problem 3-36

Considering a system of particles, and making use of notations by now familiar and self-explanatory,

$$K_{\text{rot}} = \sum_i \frac{1}{2}m_i(\vec{\omega} \times \mathbf{r}_i)^2 = \frac{1}{2}\omega^2 \sum m_i\rho_i^2 = \frac{1}{2}I\omega^2$$

.

3.19.12 Potential energy in rotational motion

Considering the rotational motion of a particle or a system of particles about an axis, one can define the potential energy of the system in a manner analogous to the potential energy in translational motion. Referring to the rotational motion of a particle for the sake of simplicity, and assuming that the field of force under which the motion of the particle takes place is a conservative one, one can define the potential energy at a point as the work done in the rotation of the particle from a chosen reference position (corresponding to a certain value of the angle (say, ϕ_0) of rotation) to the point under consideration (corresponding to angle, say, ϕ), with a negative sign.

Denoting the potential energy in rotational motion by V_{rot} , one can state the principle of conservation of energy in the context of a pure rotational motion of a system of particles

about any given axis in the form

$$K_{\text{rot}} + V_{\text{rot}} = \text{constant}. \quad (3-169)$$

3.19.13 Calculation of moments of inertia

In determining the rotational motions of systems of particles or of rigid bodies (in this context, see sec. 3.20.1) under given forces, it is often found necessary to know the moments of inertia of these systems about given axes. The calculation of the moment of inertia of a system of particles (and, in particular, of a rigid body) about any given axis relies on a number of theorems relating to moments of inertia, especially the *theorem of parallel axes*, and the *theorem of perpendicular axes* (see below). In addition, the theorem of additivity of moments of inertia (see sec. 3.19.13.4) is also employed (mostly without explicit mention of it) in determining the moments of inertia of various bodies. In what follows (sections 3.19.13.1, 3.19.13.2, 3.19.13.4), we consider a system (S) made up of N particles of masses m_1, m_2, \dots, m_N , located at any given instant of time at positions P_1, P_2, \dots, P_N , and address the question of obtaining useful information relating to the moments of inertia of this system with respect to specified axes.

3.19.13.1 The theorem of perpendicular axes

The theorem of perpendicular axes holds for a *planar* system of particles, i.e., we assume for the present that the points P_1, P_2, \dots, P_N all lie in a plane, say P . Let AB be an axis perpendicular to this plane. Suppose that two other axes A_1B_1 and A_2B_2 , perpendicular to each other, lie in the plane P , where all the three axes AB, A_1B_1, A_2B_2 are concurrent (with O as their common point, see fig. 3-33). The theorem then relates the moment of inertia (I) of the planar system S about the axis AB to the moments of inertia, say I_1 and I_2 , about the axes A_1B_1 , and A_2B_2 respectively, as

$$I = I_1 + I_2. \quad (3-170)$$

Thus, knowing any two of the three moments of inertia (I, I_1, I_2), the third can be

determined from the above relation.

Problem 3-37

Establish the relation (3-170).

Answer to Problem 3-37

Consider a Cartesian co-ordinate system with its origin at O and with x-, y-, and z-axes along A_1B_1 , A_2B_2 , and AB respectively, as shown in fig. 3-33. Let the co-ordinates of P_i with respect to these axes be x_i , y_i , $z_i = 0$ ($i = 1, \dots, N$). Then the squared distances of this point from AB, A_1B_1 , and A_2B_2 are respectively, $\rho_i^2 = x_i^2 + y_i^2$, $\rho_{i1}^2 = y_i^2$, and $\rho_{i2}^2 = x_i^2$ (note the apparent anomaly in the use of the indices 1 and 2 in ρ_{i1} and ρ_{i2}), while the moments of inertia are, by definition, $I = \sum_i m_i \rho_i^2$, $I_1 = \sum_i m_i \rho_{i2}^2$, and $I_2 = \sum_i m_i \rho_{i1}^2$.

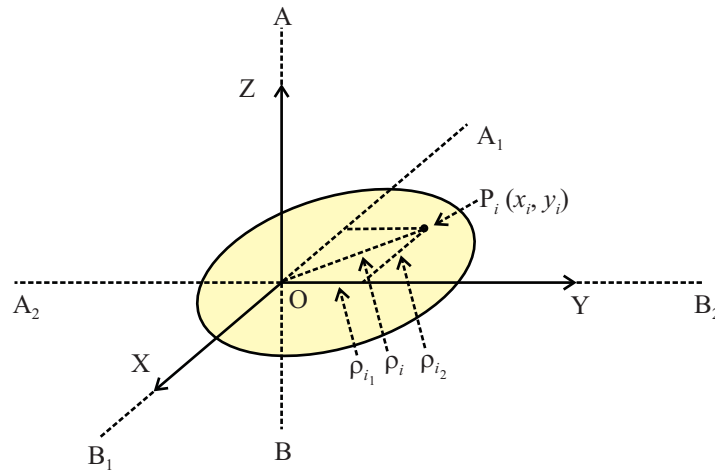


Figure 3-33: Illustrating the theorem of perpendicular axes; P_i is a point particle of mass m_i on a laminar body or a planar system of particles, having co-ordinates $x_i, y_i, z_i = 0$ with co-ordinates axes chosen as shown (the plane P mentioned in text is thus the x-y plane here); ρ_i is the distance of the point under consideration from the origin, with projections ρ_{i1} and ρ_{i2} on the y- and x-axes; the moments of inertia about the x-, y- and z-axes are related as in eq. (3-170).

The theorem of perpendicular axes is of considerable use in determining the moments of inertia of *laminar* bodies, i.e., ones in the form of planar sheets.

3.19.13.2 The theorem of parallel axes

We now remove the requirement that the points P_1, P_2, \dots, P_N corresponding to the instantaneous positions of the particles making up the system S be all co-planar, i.e., assume these points to have arbitrary locations in space. Let O be the center of mass of the system S and AB an axis passing through O . Consider any other axis, say, $A'B'$ parallel to AB . Let the moments of inertia of the system S about AB and $A'B'$ be I and I' respectively. Then the theorem of parallel axes states that the two moments of inertia are related as

$$I' = I + Md^2, \quad (3-171)$$

where $M(= \sum_i m_i)$ is the total mass of the system S , and d is the distance between the two axes AB and $A'B'$ (see fig. 3-34). Thus, knowing the moment of inertia I about an axis through the center of mass, one can determine the moment of inertia I' about any parallel axis from relation (3-171).

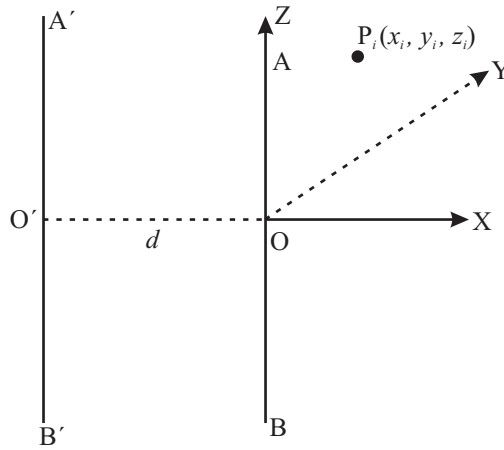


Figure 3-34: Illustrating the theorem of parallel axes; P_i is a particle belonging to the system under consideration, with mass m_i and co-ordinates (x_i, y_i, z_i) relative to a set of Cartesian co-ordinate axes chosen as shown; the axes AB and $A'B'$ are parallel, with the former passing through the center of mass of the system located at the origin O ; O' is a point on $A'B'$ lying on the x -axis; the moments of inertia about the axes AB and $A'B'$ are related as in eq. (3-171), where d stands for the distance between the two axes.

Problem 3-38

Establish the relation (3-171).

Answer to Problem 3-38

Let P be the plane passing through the center of mass O and perpendicular to the axes AB, A'B', intersected by A'B' at O'. Choose axes as in fig. 3-34, with origin at O, the x-axis along O'O, and the z-axis along OA. Considering the particle at P_i with co-ordinates, say, x_i, y_i, z_i the squared distances of the point from AB and A'B' are $\rho_i^2 = x_i^2 + y_i^2$, and $\rho_i'^2 = (x_i + d)^2 + y_i^2$. Thus $I' = \sum_i m_i \rho_i'^2 = \sum_i (m_i \rho_i^2 + m_i d^2 + 2m_i x_i d) = I + M d^2 + 2d \sum_i m_i x_i$. The sum $\sum_i m_i x_i$ is nothing but MX, where X stands for the x-co-ordinate of the center of mass, with the center of mass itself chosen as the origin. This gives eq. (3-172).

3.19.13.3 Radius of gyration

In this context, the term *radius of gyration* may be referred to. If I denotes the moment of inertia of a body of mass M about any given axis, then its radius of gyration (say, k) about that axis is defined by the relation

$$I = M k^2. \quad (3-172)$$

Thus, if one imagines a single particle of mass M , the mass of the body under consideration, to be placed at a distance from the axis equal to the radius of gyration about it, then its moment of inertia about the axis will be the same as that of the body. In other words, the radius of gyration is the distance of an *equivalent* point mass from the axis, the equivalence being in respect of the moment of inertia.

The theorem of parallel axes stated in sec. 3.19.13.2 implies the following relation between the radius of gyration of a body about any given axis and that about a parallel axis passing through the center of mass of the body. Denoting the two radii of gyration

by K and K_{cm} respectively, one has

$$K^2 = K_{\text{CM}}^2 + d^2, \quad (3-173)$$

where d stands for the distance between the two axes, and M for the mass of the body.

3.19.13.4 Additivity of moments of inertia

The definition of moment of inertia immediately implies the following:

If I_1 and I_2 be the moments of inertia of two systems of particles (or of two bodies) S_1 and S_2 about an axis AB, then the moment of inertia of the composite system made up of S_1 and S_2 about the same axis AB will be

$$I = I_1 + I_2, \quad (3-174)$$

(reason this out).

Evidently, this result can be extend to any number of systems making up a composite system.

3.19.13.5 Moments of inertia: examples

A. Moment of inertia of a thin cylindrical shell and of a circular ring.

Consider a thin cylindrical shell of mass M and of radius R . Imagining the shell to be made up of a large number of particles, each particle of mass, say, m is located at a distance R from the axis of the cylinder (fig. 3-35). Hence the moment of inertia of the shell about the axis of the cylinder can be written as $\sum mR^2 = R^2 \sum m$ where, in summing over all the constituent particles, R^2 can be taken out of the summation since it is the same for all the particles. What remains is the summation over the masses of the constituent particles, which gives, for the required moment of inertia,

$$I = MR^2. \quad (3-175)$$

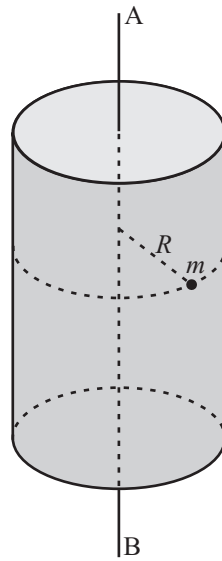


Figure 3-35: Calculating the moment of inertia of a cylindrical shell about the cylinder axis AB ; any particle of mass m on the shell is located at a distance R from the axis of the cylinder, where R stands for the radius of the shell; the moment of inertia is given by the formula (3-175).

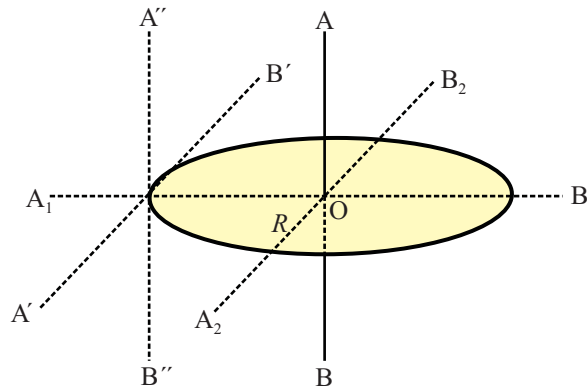


Figure 3-36: Calculating the moment of inertia of a thin ring; AB is an axis perpendicular to the plane of the ring and passing through its center; axes A_1B_1 and A_2B_2 are along two perpendicular diameters of the ring, the moments of inertia about which are equal; $A'B'$ is an axis lying in the plane of the ring, along a tangent to the ring, while $A''B''$ is an axis perpendicular to the plane of the ring, passing through a point on its rim; the moments of inertia about all these axes are related to one another by the theorems of perpendicular and parallel axes.

This result applies, in particular, to the moment of inertia of a thin ring of mass M and radius R , about an axis AB perpendicular to the plane of the ring and passing through its center (see fig. 3-36). Consider now an axis A_1B_1 lying *in the plane* of the ring and

passing through its center. The moment of inertia of the ring about such an axis is given by

$$I' = \frac{1}{2}MR^2. \quad (3-176)$$

To see why this is so, imagine a similar axis A_2B_2 (fig. 3-36) in the plane of the ring and perpendicular to A_1B_1 . Since the moments of inertia about the axes A_1B_1 , and A_2B_2 must be the same (the mass distribution in the ring being equivalent with respect to the two axes; here we assume that the mass of the ring is distributed *uniformly*), an application of the perpendicular axis theorem tells us that $I' = \frac{I}{2}$.

Consider now the axis $A'B'$ shown in fig. 3-36, which lies in the plane of the ring and is tangential to its circular periphery. The parallel axis theorem tells us that the moment of inertia about *this* axis is given by

$$I_{A'B'} = \frac{3}{2}MR^2. \quad (3-177)$$

Similarly, the moment of inertia of the ring about the axis $A''B''$ shown in fig. 3-36 is seen to be

$$I_{A''B''} = 2MR^2, \quad (3-178)$$

(check these statements out).

B. Moment of inertia of a disc.

With reference to a flat circular disc of mass M and of radius R , consider an axis AB perpendicular to the plane of the disc and passing through its center (see fig. 3-37).

In order to work out the moment of inertia of the disc about this axis, we assume the disc

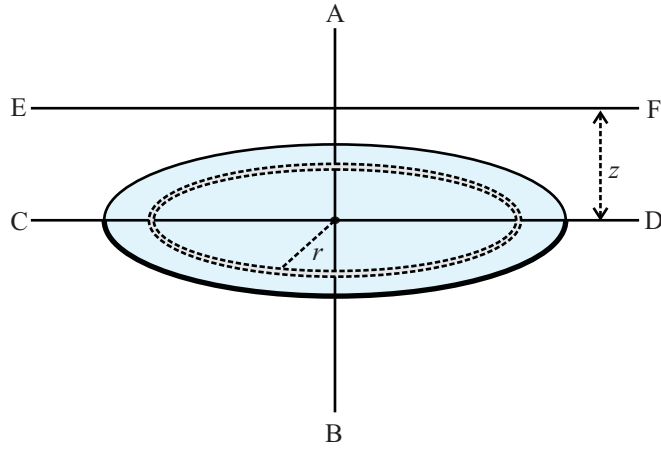


Figure 3-37: Calculating the moment of inertia of a homogeneous disc; AB is an axis perpendicular to the plane of the disc and passing through its center; CD is an axis along a diameter of the disc, while EF is parallel to the plane of the disc and intersecting AB at a distance z from the center; we take EF parallel to CD without loss of generality.

to be *homogeneous*, i.e., the mass to be distributed uniformly over its surface. One can then imagine the disc to be made up of a large number of thin rings, where a typical ring has inner and outer radii, say, r and $r + \delta r$ (see figure), the width δr being infinitesimally small. The area of the ring will then be $2\pi r \delta r$ and its mass will be $\delta m = \frac{M}{\pi R^2} 2\pi r \delta r$. In accordance with the result (3-175), the moment of inertia of the ring about the axis AB is $\delta I = \delta m r^2$. Invoking the theorem of additivity of moments of inertia, the moment of inertia of the entire disc about AB is obtained by summing up the moments of inertia of the rings making up the disc. In the limit of the width δr of the disc being made to tend to zero, one finds that the required moment of inertia is given by the integral $\int_0^R (\frac{M}{\pi R^2} 2\pi r dr) r^2$, i.e., in other words,

$$I_{AB} = \frac{1}{2} M R^2. \quad (3-179)$$

This approach of imagining a body to be made up of a large number of constituent parts and summing up the moments of inertia of these parts is a commonly employed one in the calculation of moments of inertia of bodies of various shapes about given axes.

Fig. 3-37 shows two other axes with reference to the disc, the axis CD containing a diameter of the disc, and an axis EF parallel to CD (in the plane containing AB and

CD) but at a distance z from the latter. Straightforward applications of the theorems of perpendicular axes and parallel axes give

$$I_{CD} = \frac{1}{4}MR^2, \quad I_{EF} = \frac{1}{4}MR^2 + Mz^2, \quad (3-180)$$

(check the above results out).

C. Moment of inertia of an annular ring.

Imagine a thin annular ring of mass M with inner and outer radii R_1 and R_2 . Assuming the mass of the ring to be distributed uniformly over its area, one can find its moment of inertia about an axis perpendicular to its plane and passing through the center.

Imagine a thin disc of radius R_1 that fits exactly into the central hollow of the ring so as to make up a filled disc of radius R_2 . The moment of inertia of the smaller disc is $\frac{1}{2}M_1R_1^2$ which, added to the required moment of inertia I , gives the moment of inertia of the larger disc, i.e., $\frac{1}{2}M_2R_2^2$, where $M_i = \frac{M}{R_2^2 - R_1^2}R_i^2$ ($i = 1, 2$). One thus gets

$$I = \frac{1}{2}M(R_1^2 + R_2^2). \quad (3-181)$$

Does this expression make you feel uneasy when compared with the moment of inertia of a homogeneous disc? Recall the definition of M .

D. Moment of inertia of a cylinder.

Fig. 3-38(A) depicts a homogeneous cylinder of mass M , radius R and height h , and two axes AB and CD, one along the axis of the cylinder, and the other perpendicular to it, passing through its center of mass O.

One can imagine the cylinder to be made up of a large number of thin discs, with their centers on the axis AB of the cylinder, a typical such disc being at a distance z from the center and having a thickness, say, δz , where δz may be imagined to be vanishingly small. The mass of such a disc will be $\frac{M}{h}\delta z$, and its moments of inertia about the axes

AB and CD are, by equation (3-179) and the second relation in (3-180),

$$\delta I_{AB} = \frac{1}{2} \frac{M}{h} \delta z R^2, \quad \delta I_{CD} = \frac{M}{h} \delta z \left(\frac{1}{4} R^2 + z^2 \right). \quad (3-182)$$

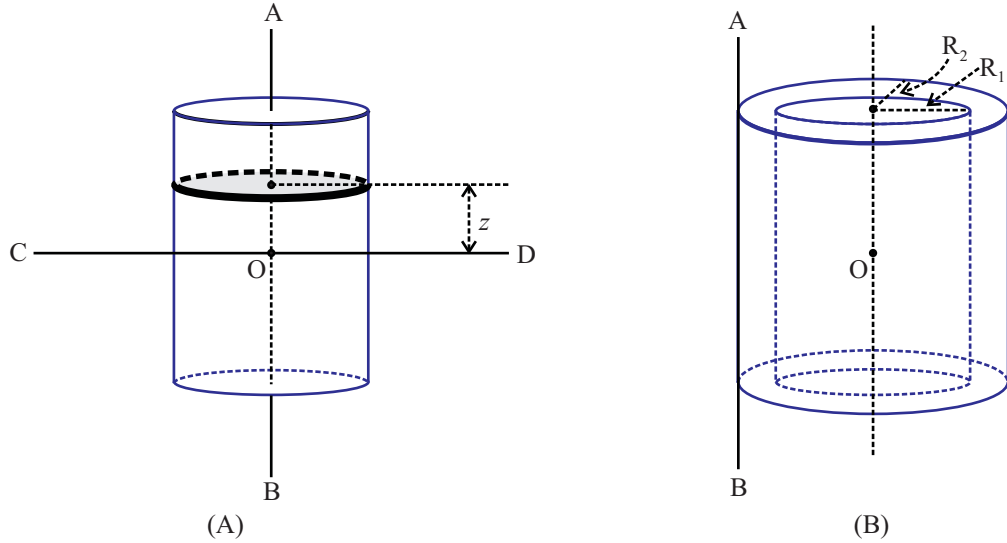


Figure 3-38: Calculating the moment of inertia of (A) a homogeneous cylinder of radius R , and (B) a hollow cylinder of inner and outer radii R_1 and R_2 ; in (A), AB is the axis of the cylinder; CD is an axis perpendicular to AB and passing through the center O; a slice in the form of a thin disc perpendicular to AB, of width δz , is shown, at a distance z from O; the moments of inertia about AB and CD are given in equations (3-183); in (B), O is the center of the cylindrical shell, and AB is an axis lying on the outer surface, parallel to the cylinder axis; the moment of inertia about AB is given by eq. (3-184).

The moment of inertia of the entire cylinder about either of these two axes is obtained by summing up the moments of inertia of all these thin discs making up the cylinder. In the limit of the thickness of each of the constituent discs going to zero, the summation reduces to an integration over z between the limits $-\frac{h}{2}$ and $\frac{h}{2}$, giving the moments of inertia

$$I_{AB} = \frac{1}{2} M R^2, \quad I_{CD} = M \left(\frac{R^2}{4} + \frac{h^2}{12} \right), \quad (3-183)$$

(check the above results out).

Problem 3-39

Problem: Moment of inertia of a hollow cylinder.

Fig. 3-38(B) shows a homogeneous hollow cylinder of inner and outer radii R_1 and R_2 and height h , along with an axis AB parallel to the cylinder axis and lying on the outer surface. Work out the moment of inertia of the cylinder about this axis.

Answer to Problem 3-39

HINT: Imagine the hollow cylinder to be made up of a large number of thin annular discs. The moment of inertia of a typical disc, of height, say δz and distance z from the center of mass O, about the axis of the cylinder is $\frac{1}{2}\delta m(R_1^2 + R_2^2)$, where $\delta m = \frac{M}{h}\delta z$ (refer to formula (3-181)). The moment of inertia about the axis AB is then, by the parallel axis theorem, $\delta m(\frac{3}{2}R_2^2 + \frac{1}{2}R_1^2)$ (parallel axis theorem). The required moment of inertia of the hollow cylinder is then obtained by summing up expressions like this, which reduces to an integration over z from $-\frac{h}{2}$ to $\frac{h}{2}$:

$$I_{AB} = \frac{1}{2}M(3R_2^2 + R_1^2). \quad (3-184)$$

E. Moment of inertia of a sphere.

Fig. 3-39 depicts a homogeneous sphere of mass M and radius R , and an axis AB passing through its center. The sphere can be imagined to be made up of a large number of thin discs, all with their planes perpendicular to AB, a typical disc being shown in the figure. Let the distance of the disc from the center of the sphere be z .

The radius of the disc is then $\sqrt{(R^2 - z^2)}$ and its mass is $\delta m = \frac{M}{\frac{4}{3}\pi R^3}\pi(R^2 - z^2)\delta z$. The moment of inertia of the disc about the axis AB is, by the relation (3-179), $\delta I = \frac{1}{2}\delta m(R^2 - z^2)$ (check this out). The moment of inertia of the entire sphere about AB is obtained by summing up the moments of inertia of all these thin discs which, in the limit of the thickness of each disc being made to go to zero, reduces to an integration over z from

$-R$ to $+R$. Thus, the moment of inertia of the sphere about the axis AB works out to

$$I_{AB} = \frac{M}{\frac{4}{3}R^3} \int_{-R}^{+R} \frac{1}{2} (R^2 - z^2)^2 dz = \frac{2}{5} MR^2. \quad (3-185)$$

An application of the theorem of parallel axes now gives the moment of inertia of the sphere about any line tangent to it, such as the axis CD in fig. 3-39:

$$I_{CD} = \frac{7}{5} MR^2. \quad (3-186)$$

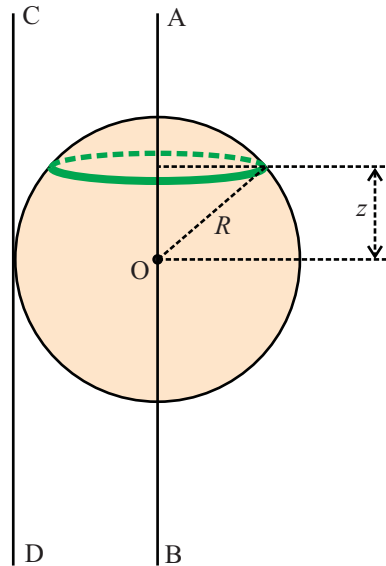


Figure 3-39: Calculating the moment of inertia of a sphere of radius R ; AB is an axis passing through the center O; a slice in the form of a thin disc of width δz is shown, at a distance z from O, the plane of the disc being perpendicular to AB; the axis CD is tangent to the sphere; the moments of inertia about AB and CD are given by equations (3-185), (3-186).

Having illustrated the method of calculation of moments of inertia by means of these examples, I append here a few other exercises for you to work out, for which refer to fig. 3-40(A)-(D).

Problem 3-40

Obtain the moment of inertia of the following bodies about axes indicated in fig. 3-40(A)-(D):

1. A thin uniform rod of mass M and length l about an axis perpendicular to its length and passing through one end.
2. A uniform rectangular lamina of mass M , length a , and breadth b , about an axis perpendicular to its plane, and passing through its center.
3. A uniform thin spherical shell of mass M and radius r about an axis along a diameter.
4. A uniform thick spherical shell of mass M and inner and outer radii a and b , about an axis along a diameter.

Answer to Problem 3-40

HINT:

1. Considering a small element of length δx at a distance x from the axis (see fig. 3-40(A)), its moment of inertia is, for vanishingly small δx , $\frac{M}{l}x^2\delta x$. Summing up all such contributions, the required moment of inertia works out to $\int_0^l \frac{M}{l}x^2 dx = \frac{Ml^2}{3}$.
2. Consider a small element of area around the point (x, y) , with the x- and y- axes parallel to the two sides of the rectangle and with the origin chosen at the center, the element itself being in the form of a tiny rectangle of sides δx , δy , both vanishingly small (fig. 3-40(B)). The moment of inertia of this element about the given axis is $\frac{M}{ab}(x^2 + y^2)\delta x\delta y$. The required moment of inertia is then $\int_{-\frac{b}{2}}^{\frac{b}{2}} dy \int_{-\frac{a}{2}}^{\frac{a}{2}} dx \frac{M}{ab}(x^2 + y^2) = \frac{M}{12}(a^2 + b^2)$.
3. Consider a rectangular co-ordinate system with the origin at the center of the shell and a small element of area on the shell around the point (x, y, z) , where the mass of the element is, say δm . Choosing the z-axis of the co-ordinate system along the given axis, the moment of inertia of the element chosen is $\delta m(x^2 + y^2)$ (see fig. 3-40(C)). From the symmetry of the problem, one can write $\int x^2 dm = \int y^2 dm = \int z^2 dm = \frac{1}{3} \int r^2 dm$ where, in each case, the integration is over the surface of the spherical shell. Thus, the required moment of inertia is $\frac{2}{3}Mr^2$.
4. The mass per unit volume of the thick shell is $\frac{M}{\frac{4}{3}\pi(b^3 - a^3)}$. Imagine the interior of the shell to be filled by a sphere of the same mass density, where the moment of inertia of this sphere about the given diameter is $\frac{2}{5} \frac{M}{\frac{4}{3}\pi(b^3 - a^3)} (\frac{4}{3}\pi a^3) a^2$ (see fig. 3-40(D), and eq. (3-185)). Adding this to the required moment of inertia of the shell, one gets the moment of inertia of a filled sphere of radius b , i.e., $\frac{2}{5} \frac{M}{\frac{4}{3}\pi(b^3 - a^3)} (\frac{4}{3}\pi b^3) b^2$. The required moment of inertia is thus $\frac{2}{5} \frac{M(b^5 - a^5)}{b^3 - a^3}$.

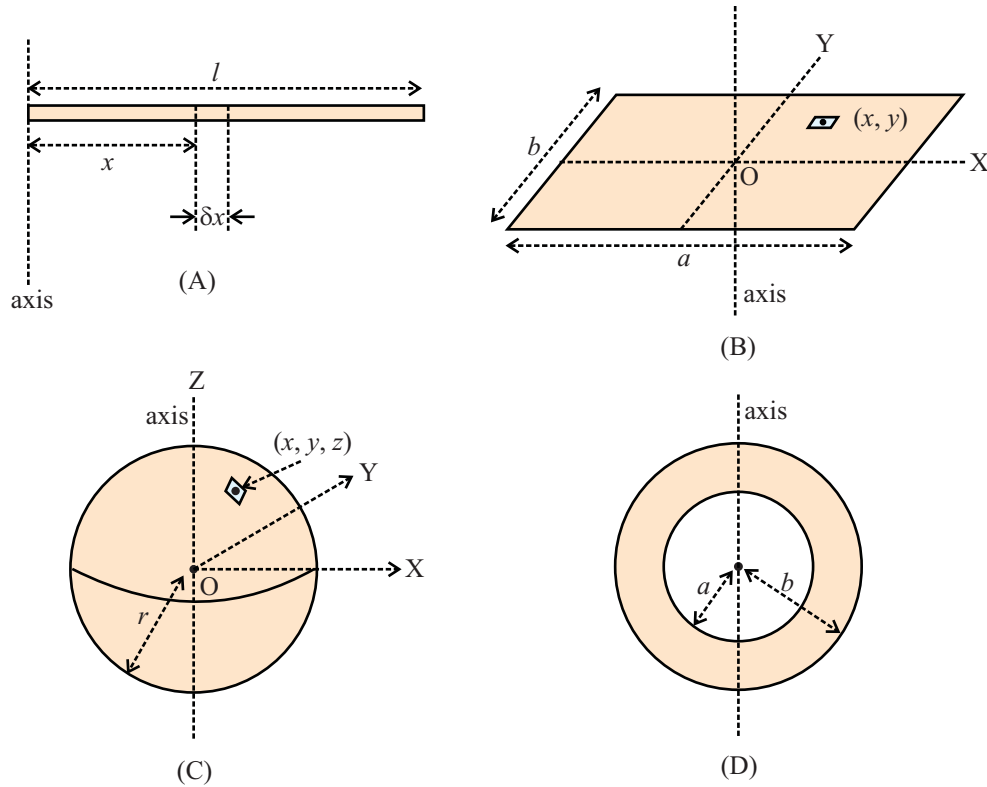


Figure 3-40: Illustrating the calculation of moments of inertia for a number of given bodies about given axes: (A) a thin rod of length l , where the axis passes through one end of the rod and is perpendicular to its length; a small element of the rod of length δx is shown, at a distance x from the axis; (B) a rectangular lamina of sides a and b , where the axis is perpendicular to the plane of the lamina and passes through its center; rectangular axes OX and OY are chosen parallel to the sides of the lamina, and a small rectangular element of sides δx , δy is chosen around the point (x, y) ; (C) a thin spherical shell of radius r , where the axis is along a diameter of the shell, chosen to be the z -axis of a rectangular co-ordinate system; a small element of area of mass δm is considered on the shell around the point (x, y, z) ; (D) a thick homogeneous spherical shell of inner and outer radii a and b , where the axis is along any diameter of the sphere; a filled sphere of radius a is imagined whose moment of inertia about the given axis, when added to the required moment of inertia of the shell, gives the moment of inertia of a filled sphere of radius b .

Problem 3-41

An annular wheel with a homogeneous mass distribution is made to rotate about its axis under an applied torque of $10 \text{ N}\cdot\text{m}$. If the mass of the wheel is 8.0 kg , and its inner and outer radii are 0.8 m and 1.1 m , what will be its angular velocity after an interval of 3.0 s from the moment it starts

revolving from rest?.

Answer to Problem 3-41

HINT: The angular acceleration of the wheel, obtained from eq. (3-163), is uniform and is given by $\alpha = \frac{N}{I}$, where $N = 10 \text{ N}\cdot\text{m}$, and $I = \frac{1}{2}M(R_1^2 + R_2^2)$ (see eq. (3-181)), with $M = 8.0 \text{ kg}$, $R_1 = 0.8 \text{ m}$, $R_2 = 1.1 \text{ m}$. The angular velocity at time $t = 3.0 \text{ s}$ after the wheel starts from rest is $\omega = \alpha t$ (see eq. (3-132)) $= 4.05 \text{ rad}\cdot\text{s}^{-1}$.

3.20 Motion of rigid bodies

3.20.1 Translational and rotational motion of rigid bodies

The concept of moment of inertia is especially useful in considerations relating to the rotational motion of *rigid bodies*. As I have already mentioned, the concept of a rigid body is a useful idealization in mechanics. The fact that the distances between the particles making up a rigid body are all fixed, i.e., independent of time, imposes strong restrictions on the possible motions of the particles. This means that the possible motions of a rigid body are relatively simple to describe. More precisely, one has the following result:

the instantaneous motion of a rigid body can be described as a translation of its center of mass along with a rotation of the body about some axis passing through the center of mass.

Fig. 3-41 illustrates schematically what this statement means. The dotted curve gives the center of mass motion of the body, the instantaneous translational velocity of the center of mass being along the tangent to this curve at any given point of time. As the center of mass moves along the dotted curve, the rigid body also gets rotated about some axis through the instantaneous center of mass, where the axis of rotation may change in course of time. For instance, when the center of mass is at O_1 , the axis of rotation is O_1M_1 , while at O_2 and O_3 , it is O_2M_2 and O_3M_3 respectively.

1. I do not enter here into a formal proof of the above statement, which constitutes the central theorem in rigid body motion.

2. There exists an alternative characterization of the instantaneous motion of a rigid body: it is a *screw motion* about some axis not necessarily passing through the center of mass. This means that the rigid body advances along the axis while at the same time undergoing a rotation about it. The screw axis need not be a fixed line but may change with time.

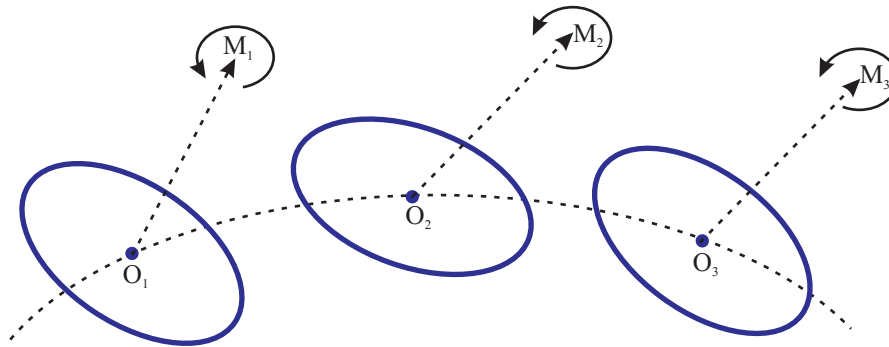


Figure 3-41: Illustrating the general nature of the motion of a rigid body; the center of mass motion takes place along the dotted curve, while there occurs simultaneously a rotation about some axis (dotted arrow) passing through the center of mass ; three positions of the body are shown along with the corresponding axes of rotation.

Thus a complete determination of the motion of a rigid body requires a solution of *two* problems: a determination of the center of mass motion, and a determination of the rotational motion about the center of mass. This means that one needs to determine at any given instant the velocity \mathbf{V} of the center of mass, along with the angular velocity $\vec{\omega}$, or equivalently, the angular momentum \mathbf{L} about the center of mass. A knowledge of the angular velocity also gives the instantaneous axis of rotation of the rigid body.

These two motions of the rigid body can be determined from equations (3-93b) and (3-149) respectively. Thus, the forces acting on a rigid body affect its motion in two ways: by changing its state of translational motion, and by changing its state of rotation. Accordingly, a system of forces acting on a rigid body is said to be in equilibrium, if it produces none of these changes in the body. This question of a system of forces acting on a rigid body being in equilibrium will be addressed in sec. 3.22.

Corresponding to the decomposition of the instantaneous motion of a rigid body into a

translation of the center of mass and a rotation about the center of mass, one can derive a simple expression of the instantaneous *kinetic energy* of the rigid body:

$$K = \frac{1}{2}MV^2 + \frac{1}{2}I\omega^2. \quad (3-187a)$$

Here M stands for the mass of the rigid body, and I denotes the moment of inertia about the instantaneous axis of rotation. The first term on the right hand side denotes the kinetic energy associated with the translational motion of the center of mass, while the second term denotes the *rotational* kinetic energy. Notice that there is a similarity between the rotational and translational terms, the moment of inertia being analogous to mass, and the angular velocity to the translational velocity. An alternative expression for the kinetic energy is

$$K = \frac{P^2}{2M} + \frac{L^2}{2I}, \quad (3-187b)$$

where $\mathbf{P} = M\mathbf{V}$ stands for the linear momentum of a particle of mass M imagined to be moving with the center of mass, and \mathbf{L} denotes the instantaneous angular momentum.

Problem 3-42

A thin homogeneous rod of length 2 m and mass 0.5 kg is held vertically with one end on the ground and then let fall. Find its kinetic energy when the rod falls to the ground, and the speed of the center of the rod at that instant, assuming that the lower end does not slip on the ground ($g = 9.8 \text{ m}\cdot\text{s}^{-2}$).

Answer to Problem 3-42

HINT: The potential energy of the rod in its initial vertical position can be worked out by imagining it to be divided into a large number of small elements. Considering a typical element of vanishingly small length δz at a height z , its potential energy is $\frac{M}{l}gz\delta z$ (M = mass of the rod, l = length of the rod). Summing up over all the elements, the potential energy of the rod is seen to be $\int_0^l \frac{M}{l}gzdz = \frac{Mgl}{2}$, this being the potential energy of an equivalent point mass M at the center of mass of the rod.

As the rod falls to the ground, its potential energy gets converted to kinetic energy, which is thus seen to be $K = 0.5 \times 9.8 \times 1.0$ J. Since the motion of the rod is one of rotation about the end touching the ground, one has $K = \frac{1}{2} I \omega^2$, where $I = \frac{Ml^2}{3}$ (refer to problem with hint in sec. 3.19.13.5, fig. 3-40(A)), and $\omega = \frac{v}{\frac{l}{2}}$, where v is the required velocity (reason this out), which thereby works out to $v = \sqrt{\frac{3gl}{4}} = 3.83 \text{ m}\cdot\text{s}^{-1}$ (approx).

The same result can be derived by referring to the equation of motion of the rod, noting that the motion is one of rotation about a horizontal axis passing through the point of contact of the rod on the ground. The relevant equation of motion is then eq. (3-163), where we note that it is nothing but the equation obtained from Newton's second law of motion in the context of rotational motion. Considering any instant of time when the inclination of the rod is θ to the vertical, the torque about the point of contact, arising from its weight Mg is $Mg \frac{l}{2} \sin \theta$ where $M = 0.5$ kg and $l = 2$ m are the mass and the length of the rod. Noting further that the moment of inertia about the axis of rotation is $\frac{1}{3} Ml^2$, the equation of motion is

$$\frac{1}{3} Ml^2 \frac{d^2 \theta}{dt^2} = \frac{1}{2} Mgl \sin \theta.$$

Multiplying both sides with $2 \frac{d\theta}{dt}$ and integrating from $\theta = 0$ to $\theta = \frac{\pi}{2}$, one obtains the velocity of the center at the time of the rod falling to ground as $v = \sqrt{\frac{3gl}{4}}$, the same result as the one obtained from the energy principle (note that the mass gets canceled in the expression for v).

NOTE: The principle of conservation of energy is thus seen to be equivalent to the result of performing an integration of the equation of motion of the system under consideration. A complete description of the motion, if necessary, can be obtained by performing a second integration on this result. .

3.20.2 Rolling motion

Mention may now be made of the *rolling motion* of a rigid body on another rigid surface. Though the concept of a rigid body is an idealization, the idea of rolling motion is still a useful one since it gives a simple and sufficiently accurate description of a certain class of motions commonly observed in our everyday experience as also in engineering practice.

The most common example of rolling motion is the motion of a *wheel* on rough ground.

On slippery ground, the wheel, in addition to rolling on the ground, may also *slide* or drag over it. In contrast to *pure* rolling, such a motion is described as one of rolling with sliding. Unless otherwise specified, the term rolling will be used to mean pure rolling.

In defining rolling motion of a rigid body A over a rigid surface B we assume that A is in contact with B at a single point, as in the rolling of a spherical body over a flat surface, or along a single line as in the rolling motion of a cylindrical roller on the ground. Pure rolling (or, simply, rolling) is then defined as a motion in which the point of contact or the line of contact of the rolling body A is instantaneously at rest on the surface B, i.e., the instantaneous velocity of the point of contact of A, or of the points of A making up the line of contact, is zero.

This implies that the instantaneous motion of A must be one of *rotation* about an axis passing through the point of contact or, in the case of the contact being along a line, a rotation about the line of contact itself. In the case of rolling with sliding there takes place, along with a rotation about such an axis, a translational motion of the point of contact or of the line of contact along the surface B.

Fig 3-42 shows a cylindrical body (A) rolling over a flat surface (S) where the instantaneous line of contact BC is along a line parallel to the axis of the cylinder and lying on its surface (such a line is referred to as a *generator* of the cylindrical surface). The instantaneous motion of the cylinder is then a rotation about this line with a certain angular velocity, say, ω in either of the two possible directions of rotation.

As the cylinder rolls on, new lines of contact are established at successive instants, and the angular velocity may change in the course of time. It may even happen that, at a later instant of time, the motion ceases to be one of pure rolling.

Referring to fig. 3-42, if O be a point on the axis of the cylinder at any given instant of time (one can conveniently choose this as the center of mass of the cylinder, whose mass distribution we assume to be homogeneous) when the instantaneous axis of rotation (i.e., line of contact) is BC, and if ON be the perpendicular dropped from O on BC, then

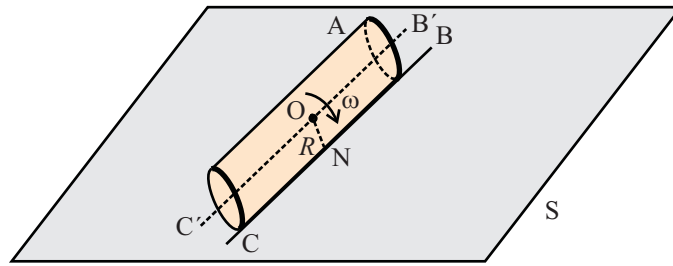


Figure 3-42: Illustrating the rolling motion of a cylinder (A) of radius R on a flat surface (S); BC is the instantaneous line of contact, where the instantaneous motion of the cylinder is one of rotation about BC with an angular velocity ω ; as the cylinder rolls on, the line of contact keeps changing.

the instantaneous motion of O is one of rotation about the axis BC, the center of rotation being N. The instantaneous velocity of O is then (see eq. (3-160)) $v = \omega R$, where R stands for the radius of the cylinder.

Continuing to refer to the example of the rolling cylinder, the motion can equivalently be described as a rotation about the axis (B'C') (parallel to BC and passing through the center of mass O) of the cylinder with the same angular velocity ω *along with a translational motion* of every point on this axis (and, in particular, of the center of mass) with a velocity ωR . This is a particular instance of the result that the instantaneous motion of a rigid body can be described as a motion of the center of mass along with a rotation of the body about some axis passing through the center of mass.

As mentioned in sec. 3.20.1, another convenient characterization of the instantaneous motion of a rigid body is to describe it as a screw motion. In the case of pure rolling, the screw axis is the instantaneous axis of rotation passing through the point of contact. The screw motion in this case is simple in nature, namely, one where the velocity along the screw axis is zero. Indeed, any rotational motion of a rigid body is of this special type.

Problem 3-43

A homogeneous rigid cylindrical body of mass M and radius R rolls without sliding on flat ground

with a velocity v . Work out the kinetic energy of the cylinder.

Answer to Problem 3-43

Making use of equations (3-187a), (3-183), and (3-160), one obtains

$$K = \frac{1}{2}Mv^2 + \frac{1}{2}\left(\frac{1}{2}MR^2\right)\frac{v^2}{R^2} = \frac{3}{4}Mv^2. \quad (3-188)$$

The same result can be derived by recalling that the motion of the cylinder is one of pure rotation about the instantaneous line of contact with an angular velocity $\frac{v}{R}$. Try this out.

If the motion of the cylinder is one of rolling with sliding then the velocity (v) of the center of mass and the angular velocity of the center of mass will not be related as $v = a\omega$. The difference between the two will then denote the velocity of the instantaneous line of contact.

The motion of pure rolling of a cylindrical body is relatively simple compared to the rolling motion of, say, a spherical body on a flat surface, where there is a single point of contact instead of a line of contact at any given instant of time (fig.3-43). The instantaneous axis of rotation in this case can be any line passing through the point of contact (A in the figure). As can be seen from the figure, the instantaneous angular velocity can be resolved into two components, one along the plane S on which the rolling occurs, and the other along AO, i.e., the line joining the point of contact and the center of the sphere. The latter corresponds to a *spinning* motion of the sphere about AO, in which the instantaneous velocity of the center of mass (O, the sphere being assumed to be homogeneous one for the sake of simplicity) is zero. Along with this spinning motion, the sphere possesses an instantaneous rotational motion about some axis, say, CAB lying in the plane S.

If ω be the angular velocity of the rotation about the axis CAB, then the instantaneous velocity of the center of the sphere is given by $v = a\omega$, where a stands for the radius of the sphere.

The rolling motion of a wheel in the shape of a circular disc on a flat plane is essentially

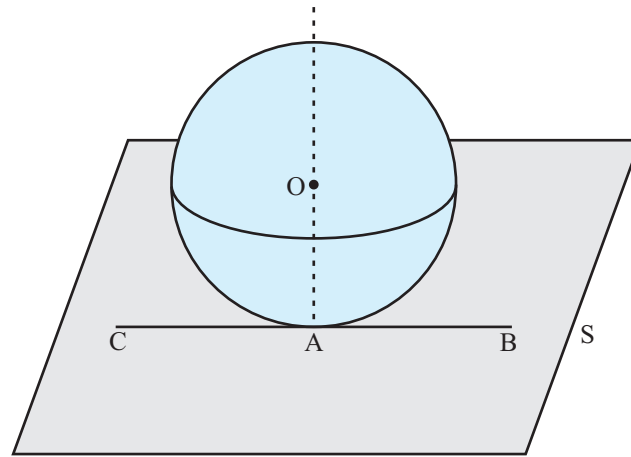


Figure 3-43: Pure rolling of a sphere (center O) on a plane (S); A is the instantaneous point of contact; the instantaneous motion of the sphere is one of rotation about some axis passing through A ; this can be resolved into a spinning motion about AO , and a rotation about an axis such as CAB in the plane S .

similar to that of a spherical body. Fig. 3-44 shows a wheel where the plane of the wheel is assumed to be vertical for the sake of simplicity. In a rolling motion of the wheel, the instantaneous axis of rotation passes through the point of contact (A), and the angular velocity can have a component (say, ω_1) along the vertical line AO as well as along a line lying in the plane S (which we assume to be horizontal) on which the wheel rolls. The angular velocity about the latter can, in turn, be resolved into a component (say, ω_2) about a line AB lying in the plane of the wheel (assumed vertical for the sake of simplicity) and one (ω) about a line AC perpendicular to its plane.

Of these three, ω_1 represents a spinning motion in which the velocity of the center of the wheel (its center of mass if the mass distribution in the wheel is assumed to be a uniform one) is zero, while ω_2 represents a tilting motion in which the plane of the wheel tends to tilt away from the vertical while the instantaneous point of contact does not change. Finally, ω represents a motion we commonly refer to as rolling, in which the wheel rolls over to a new point of contact at each instant.

In general, the rolling motion involves all three components whereby the wheel can spin around, tilt, and roll forward. A combination of the three enables the wheel to turn

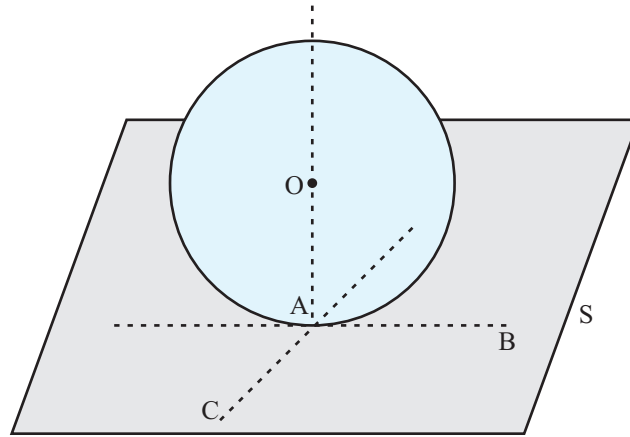


Figure 3-44: Pure rolling of a wheel; the plane of the wheel is assumed to be vertical for the sake of simplicity; O is the center of the wheel while A is the instantaneous point of contact; the instantaneous motion of the wheel is made up of a spinning motion about AO, a rotation about the axis AB lying in S, which makes the plane of the wheel tilt away from the vertical, and a rotation about the axis AC, which makes the wheel roll on.

around as it rolls. In the simplest case in which the plane of the wheel is vertical and the instantaneous axis of rotation lies in the horizontal plane S along the line AC perpendicular to the plane of the wheel, there is no spinning, tilting, or turning motion, and the wheel rolls with its plane remaining vertical. The velocity of the center of the wheel in this case is

$$v = \omega a, \quad (3-189)$$

where a stands for the radius of the wheel.

The relation (3-189) gives the criterion for pure rolling of the wheel with its plane vertical. A deviation from this equality in this case corresponds to a motion involving sliding (in which the instantaneous point of contact drags along the plane S) along with the rolling motion.

3.20.3 Precession

3.20.3.1 Precession under an applied torque

Fig. 3-45 depicts a rigid body of symmetrical shape (say, in the form of a ring) rotating about its axis of symmetry (dotted vertical line in the figure; say, the z -axis in a Cartesian co-ordinate system) with its center of mass held fixed. We assume the angular velocity of rotation to be sufficiently large.

As we have seen (sec. 3.20.1), the general motion of a rigid body is made up of a translation of its center of mass along with a rotation about an axis passing through the center of mass. The translational motion is absent in the present instance.

Suppose now that a torque is made to act on the body by means of a pair of equal and opposite forces applied as shown, where the magnitude and direction of the forces are held fixed with reference to the body. As seen in the figure, the torque on the body acts along a direction perpendicular to its instantaneous axis of rotation (i.e., in the present instance, the x -axis of the Cartesian co-ordinate system chosen).

It is then found that, instead of the axis of the body getting tilted in the y - z plane as might be naively expected, the body executes a slow turning motion about the y -axis while continuing to rotate about its own axis of symmetry. The latter remains in the z - x plane and turns about the y -axis at a constant rate. For a given magnitude of the torque (say, M) applied to the body, the angular velocity (Ω) of turning of the axis of rotation is found to be in inverse proportion to the angular momentum (L) of rotation about the symmetry axis. Such a turning motion of the axis of rotation is referred to as *precession*.

The explanation of such precessional motion is obtained from the basic equation (3-158) describing the angular motion of a system of particles. Additionally, we make use of a result in the theory of rotational motion that states that, for a body rotating about a symmetry axis, the angular momentum is directed along the axis of rotation. Moreover,

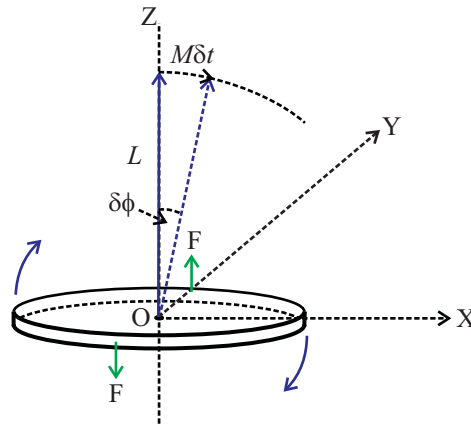


Figure 3-45: Illustrating the idea of precession of a rotating body; the body is assumed to be in the form of a ring with its symmetry axis along OZ at any particular instant; forces of magnitude F act at the two ends of a diameter in opposite directions to make up a torque directed along OX; as a result of this torque, the body precesses about the axis OY, with the plane of the ring slowly turning about OY as shown by the curved arrows; the angular momentum vector (of magnitude L), and the impulse of the torque (magnitude $M\delta t$) over a short time interval δt are indicated; the changed direction of the angular momentum vector at the end of this interval is shown by a dotted arrow; the tip of the angular momentum vector traces out a circle in the z-x plane.

if the angular velocity of rotation is large enough, then the motion of the body continues to be, in the main, a rotation about the symmetry axis in spite of the applied torque, i.e., in other words, the instantaneous angular momentum of the body continues to be directed along the symmetry axis.

In the figure, the angular momentum vector is shown at the instant under consideration when the axis of rotation points along the z-axis of a Cartesian co-ordinate system, while the direction of the torque is along the x-axis. The short arrow parallel to the x-axis represents the impulse $M\delta t$ of the torque in a short time interval δt and, according to eq. (3-158) gives the change in the angular momentum in this time interval. The dotted arrow slightly inclined to the vertical then depicts the angular momentum at the end of the interval δt . Evidently, assuming δt to be sufficiently small, the angular momentum remains unchanged in magnitude while getting rotated in the z-x plane by an angle $\delta\phi = \frac{M}{L}\delta t$.

This means that the axis of rotation of the body, which in the present instance is its symmetry axis, gets rotated in the z-x plane about the y-axis at a rate $\Omega = \frac{\delta\phi}{\delta t} = \frac{M}{L}$. At

the end of the interval δt , there occurs a similar turning of the axis of rotation in the z-x plane in a subsequent interval since, by assumption, the applied torque remains perpendicular to the axis of rotation, which means that the entire figure simply gets rotated by $\delta\phi$ about the y-axis.

This slow turning of the rotation axis of the body at a rate $\Omega(= \frac{M}{L})$ due to the applied torque of magnitude M is precisely what constitutes the precession of the body under consideration.

3.20.3.2 Precession of a heavy top

The example of precession considered above corresponds to the special case in which the symmetry axis maintains a right angle with the axis about which the precession takes place (i.e., the y-axis in fig. 3-45; the symmetry axis turns in the z-x plane). Fig. 3-46 depicts the more general case of precession of a *heavy top* set spinning about its symmetry axis OT with a large value of the angular velocity (ω) which slowly turns about the vertical axis OZ, this slow turning constituting the precession of the top.

The torque acting on the top is made up of the pull of gravity acting through its center of mass C, and an equal and opposite reaction force exerted by the ground at the point of contact O. The torque acts along AO, perpendicular to both the instantaneous position of the symmetry axis and the vertical direction OZ, the angle between the latter two being, say, α , representing the tilt of the top.

Considering the directed line segments representing the angular momentum vector (of magnitude L) at any given instant along OT and the impulse of the torque of magnitude $M\delta t$ parallel to AO, it is seen that the angular momentum vector turns through an angle $\delta\phi = \frac{M\delta t}{L \sin \alpha}$ (check this out!), corresponding to which the angular velocity of precession is

$$\Omega = \frac{M}{L \sin \alpha}. \quad (3-190)$$

This reduces to the value $\frac{M}{L}$ in the special case $\alpha = \frac{\pi}{2}$.

For the situation depicted in fig. 3-46, the top precesses with the point O fixed on the ground by means of frictional forces exerted by the latter (see sec. 3.23 for a brief introduction to static and dynamic friction). The center of mass motion of the top consists in this case of a uniform rotation in a circular path with angular velocity Ω . The centripetal acceleration for this motion is provided by the frictional force acting on the top at O. If, however, the top is set spinning on smooth ground, the frictional force does not come into play, and the top will then precess with its center of mass C stationary, i.e., the axis of precession will now be along the vertical line passing through C.

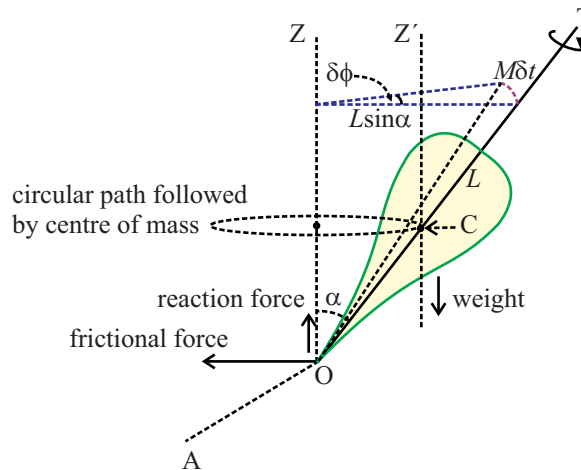


Figure 3-46: Precession of a heavy symmetric top tilted by an angle α from the vertical; the instantaneous position of the spinning axis is OT, along which points the angular momentum vector of magnitude L ; the torque made up of the weight of the top and the normal reaction of the ground, acts along AO (pointing into the plane of the figure), perpendicular to both the symmetry axis and the vertical direction OZ; as a result of the impulse of the torque ($M\delta t$) acting over a short time interval δt , the angular momentum vector turns about OZ by an angle $\delta\phi$; the turning motion continues, constituting the precession; the center of mass moves along a circular path, the necessary centripetal acceleration being provided by the force of friction at A; the direction of precession is related to the vertically upward direction in the right handed sense.

3.20.3.3 Precession of the earth's axis of rotation

The earth may be looked upon as a rigid body revolving around the sun and at the same time spinning about its own axis, the axis being inclined to the plane of the earth's orbit, i.e., the plane in which the motion around the sun takes place (fig. 3-47).

While the sun exerts a gravitational pull on the earth, making it move along a nearly circular orbit (see sec. 5.5), at the same time, it exerts a *torque* on the earth, where a similar torque on the earth is exerted by the moon as well. The two torques in combination act in a direction tending to pull the axis of rotation of the earth toward the perpendicular to the plane of the earth's orbit. In other words, the gravitational effect of the sun and the moon on the earth can be described in terms of a force through its center of mass, and a couple (see sec. 3.22.3.3), where the plane of the couple (a plane perpendicular to the direction of the moment of the couple, i.e., the net torque due to it) contains the force.

One can, in an approximate sense, combine the couple and the force into a single force, but then the latter would not pass through the center of mass of the earth.

This is shown in fig. 3-47 in which the angular momentum vector (of magnitude L) of the earth, which is almost identical with the angular momentum due to the spinning motion about its own axis, is shown at a given time, and the impulse of the torque (of magnitude, say, M) in a short time interval δt is also indicated. As a result of the torque, the angular momentum vector gets turned about the axis OZ, constituting the precession of the rotation axis, the angular velocity of precession being once again given by $\Omega = \frac{M}{L \sin \alpha}$ where α , the angle between the direction of the pole star and the perpendicular to the plane of the ecliptic, has a value ≈ 0.41 . While the spinning motion occurs in an anticlockwise sense about the earth's axis, the precession occurs in a clockwise sense about the direction perpendicular to the plane of the earth's orbit (when viewed from the direction of the pole star). This precession is commonly referred to as the 'precession of the equinoxes', and is characterized by a time period $T = \frac{2\pi}{\Omega} \approx 26000$ years. However, strictly speaking, the precession is not a regular one, having other perturbations mixed with it.

The torque exerted by the sun on the earth, as also that exerted by the moon, arises because of a deviation of the shape of the earth from a spherical one. The earth can

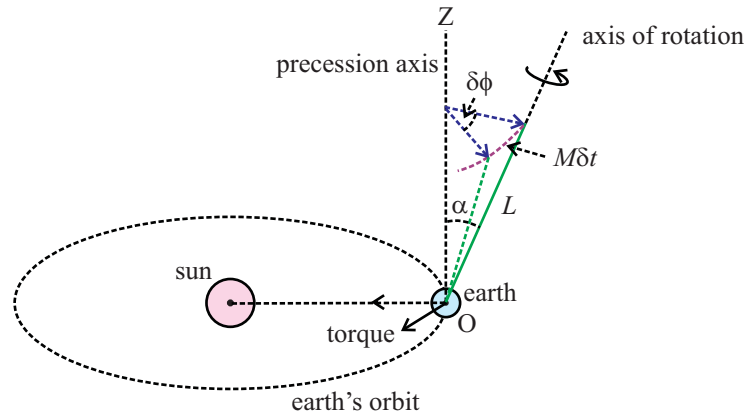


Figure 3-47: Precession of the earth's axis of rotation; the instantaneous angular momentum of the earth at position O points along the tilted axis of the earth, inclined at an angle α to OZ, the perpendicular to the plane of the earth's orbit at O; the direction of the torque due to the sun and the moon is shown, giving the impulse ($M\delta t$) over a short interval δt ; the angular momentum vector turns about OZ to the new position shown by the dotted line.

be described approximately as an *oblate spheroid*, with the equatorial region bulging out and the polar diameter compressed by comparison. When one considers this non-spherical shape along with the tilt of the earth's axis, and takes into account the fact that the gravitational pull of the sun on the various different volume elements making up the earth differ from one another owing to the difference of the distances of these elements from the sun (in this context, refer to sec. 5.5.5 where the *tidal force* is introduced), one ends up with a net force through the center of mass of the earth *and* a torque in a direction as described above.

3.20.3.4 Free precession

A different kind of precession is exhibited by a rotating body in *torque-free motion* if the body is set spinning about an axis *other than its symmetry axis*. The angular momentum vector of the body remains unchanged in magnitude and direction (since there is no torque on the body), but the body itself precesses about the direction of this vector, with the symmetry axis describing a cone about it.

Corresponding to any given point in a rigid body, there can be found three *principal axes* of inertia, where a principal axis is characterized by the fact that, in the case

of rotation about this axis, the angular momentum points along the axis itself. For a body with an axis of symmetry passing through the point under consideration, the symmetry axis happens to be one of the three principal axes of inertia. If the moment of inertia of the body about the symmetry axis be less than or equal to each of the moments of inertia about the other two principal axes, then the body can perform stable rotations about the symmetry axis, and a deviation of the axis of rotation from the symmetry axis leads to a stable motion of the body involving precession of the axis of rotation about the symmetry axis.

3.21 Rotating frames of reference

Let us consider an inertial frame of reference (S) with respect to which a second frame (S') possesses a *rotational motion*. Fig. 3-48 shows two Cartesian co-ordinate systems in S and S' where, for the sake of simplicity, we assume that the rotational motion of S' takes place about the z-axis (OZ) of S. Moreover, the z-axis of the co-ordinate system in S' is chosen to be coincident with the z-axis of the system in S. We shall, for the sake of brevity, use the symbols S and S' to denote the two co-ordinate systems as well as the corresponding frames of reference.

Imagine a particle of mass m at the position P at any given instant of time. As indicated in sec. 3.10.3, the equation of motion of the particle in S', which is a *non-inertial* frame because of its rotational motion, is of the form of eq. (3-52), where \mathbf{F} is the 'real' force acting on it, due to its interaction with other particles or bodies, while \mathbf{G} denotes the *pseudo-force*, or inertial force, arising by virtue of this rotational motion of S'.

The pseudo-force \mathbf{G} in the rotating frame S' is commonly expressed in the form

$$\mathbf{G} = \mathbf{G}_1 + \mathbf{G}_2. \quad (3-191)$$

Of the two, \mathbf{G}_1 depends only on the instantaneous position (\mathbf{r}) of the particle under consideration while \mathbf{G}_2 depends on its instantaneous velocity (\mathbf{v}), where both depend, in

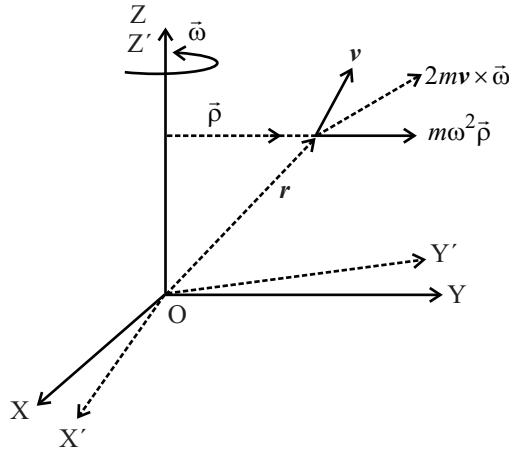


Figure 3-48: Illustrating centrifugal and Coriolis forces in a rotating frame; $OXYZ$ is a Cartesian co-ordinate system in an inertial frame (S) while $OX'Y'Z'$ is a co-ordinate system in a frame (S') rotating about OZ with an angular velocity ω ; v denotes the instantaneous velocity of a particle located at P ; PN is the perpendicular dropped from P on OZ ; $\vec{\rho}$ is the vector extending from N to P ; the centrifugal force (eq. (3-193)) and the Coriolis force (eq. (3-194)) experienced by the particle in the frame S' are indicated schematically.

addition, on the angular velocity ($\vec{\omega}$) of S' relative to S . These two forces appearing in S' are referred to as, respectively, *centrifugal* and *Coriolis* forces.

The angular velocity $\vec{\omega}$ is defined by looking at the co-ordinate system in S' as a rigid body (made up of the three co-ordinate axes rigidly attached to one another) or a set of points on the co-ordinate axes rigidly attached to one another. Since this set of points possesses a rotational motion relative to S , its angular velocity ($\vec{\omega}$) can be defined as in sec. 3.19.10. For the sake of simplicity, we assume that $\vec{\omega}$ is constant in time. In the case of a non-uniformly rotating frame the expression for the pseudo-force involves a term other than G_1 and G_2 of eq. (3-191).

Imagine a perpendicular dropped from P , the instantaneous position of the particle under consideration, on the axis of rotation of the frame S' , which we have assumed to be the z -axis of S , as also that of S' . The instantaneous distance of P from the axis of rotation is then

$$\rho = \sqrt{x^2 + y^2}, \quad (3-192)$$

where x, y denote the x- and y- co-ordinates of P in S' respectively. If the foot of the perpendicular dropped from P on the rotation axis be N, then the vector extending from N to P is of magnitude ρ , and we denote this vector as $\vec{\rho}$. The centrifugal force is then given by the expression

$$\mathbf{G}_1 = m\omega^2\vec{\rho}. \quad (3-193)$$

In other words, the centrifugal force is directed away from the axis of rotation, is proportional to the distance of the particle from this axis, and is proportional to the square of the angular velocity of the rotating frame.

The Coriolis force, on the other hand, is given by the expression

$$\mathbf{G}_2 = 2m\mathbf{v} \times \vec{\omega}, \quad (3-194)$$

where \mathbf{v} stands for the instantaneous velocity of the particle under consideration.

Thus, for instance, the Coriolis force vanishes for a particle moving in a direction parallel to the z-axis in fig. 3-48, while the centrifugal force vanishes if the particle is located on the z-axis.

Note that each of the two pseudo-forces is proportional to m , the mass of the particle under consideration, as is the pseudo-force given by eq. (3-53) arising in the case of an accelerated translational motion of the frame S' relative to S . The real force on a particle, on the other hand, is not, in general proportional to its mass (the great exception being the force of gravitational interaction, see chapter 5).

The fact that the pseudo-force acting on a particle is proportional to its mass and, other than the mass, position, and velocity of the particle, depends only on the acceleration of the frame of reference with respect to any inertial frame, may be made use of in working out the resultant of the pseudo-forces acting on a *system* of particles. For instance, *the resultant of the inertial forces in a non-rotating accelerated frame, acting on all the particles taken together, passes through the center of mass of the system.*

Effects of inertial forces are observed in numerous man-made set-ups and natural phenomena. The outward pull experienced by a child in a merry-go-round is due to the centrifugal force acting on her in a frame rotating with the merry-go-round. The rotation of air currents in cyclones is caused by the Coriolis force in the earth-bound frame which arises because of the fact the latter is a rotating non-inertial frame. It is found that, because of the Coriolis force in an earth-bound frame, the right bank of a river in the northern hemisphere gets broken and eroded quicker than the left bank. In the case of a river in the southern hemisphere, on the other hand, the left bank gets broken at a faster rate (Ferrel's law). All these facts as also a number of other observations can be explained in terms of the centrifugal and Coriolis forces arising in rotating frames of reference.

The following problem (see fig. 3-49) illustrates the idea of centrifugal force in a rotating frame in the context of a bead mounted on a rotating hoop.

Problem 3-44

Consider a bead mounted on a rotating circular hoop (fig. 3-49) on which it can slide without friction, where the plane of the hoop at any instant of time is vertical, the axis of rotation being along a vertical line passing through its center. If the angular velocity of the hoop be ω , and its radius be R , obtain the position of equilibrium of the bead on the hoop.

Answer to Problem 3-44

Consider a frame of reference rotating with an angular velocity ω , in which the hoop appears stationary. Looking at fig. 3-49, the forces acting on the bead B are its weight mg acting vertically downward (g = acceleration due to gravity), the force of reaction N acting along BO, where O is the center of the hoop, and the centrifugal force $F = m\omega^2\rho$, acting horizontally away from the axis of rotation where ρ (=AB) is the distance of the bead from the axis of rotation PO. If the bead be in equilibrium, the sum of the resolved components of the forces along any given direction has to be zero (see sec. 3.22.1.1 for basic ideas relating to the condition of equilibrium of a set of concurrent forces). Taking this direction to be along the tangent at B, one has $mg \sin \theta = F \cos \theta$, where θ is the angle between OB and OA. From the figure, $\rho = R \sin \theta$, and thus, $\cos \theta = \frac{g}{\omega^2 R}$. The bead will rest at the lowest point P on the hoop if the angular velocity is less than $\sqrt{\frac{g}{R}}$, in which case $F = 0$

and the equilibrium of the bead requires $N = W$. If, on the other hand, $\omega > \sqrt{\frac{g}{R}}$, the position of equilibrium is given by $\theta = \cos^{-1} \frac{g}{\omega^2 R}$.

Incidentally, the situation depicted in fig. 3-49 differs from that in fig. 3-32(B) in that, in the former, the hoop is made to rotate about its vertical axis at a *constant* rate while, in the latter, the angular velocity of rotation of the hoop is variable, depending on the sliding motion of the bead.

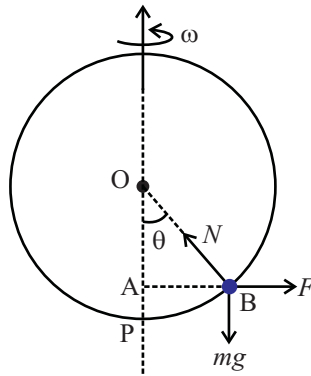


Figure 3-49: A bead (B) capable of sliding on a smooth circular hoop rotating about a vertical axis OP where O is the center of the hoop, whose plane at any instant of time is vertical; the normal reaction of the hoop on the bead acts along BO while the weight of the bead acts vertically downward; in a rotating frame, the centrifugal force on the hoop acts horizontally away from the axis OP ; for small values of ω the bead rests at the lowest point P of the hoop; at higher values of ω the bead slides to a new equilibrium position such as B .

3.22 Reduction of a system of forces

3.22.1 Introduction

We now consider a system of forces acting on a particle or on a rigid body. A system of forces acting on a particle is distinguished by the fact that all the forces in the system are *concurrent* since they all act through the same point. The forces acting on a rigid body, however, need not be concurrent. The question we will now address is the following: can the system of forces under consideration be reduced to one that is relatively *simple* to describe and has the *same* effect on the state of motion of the particle or the rigid body as does the given system?

3.22.1.1 Concurrent forces

In the case of a concurrent system of forces, this question has already been answered in principle in sec. 3.5.5: a concurrent system made up of forces $\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_N$ can be reduced to a single force \mathbf{F} ,

$$\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2 + \dots + \mathbf{F}_N = \sum_{i=1}^N \mathbf{F}_i, \quad (3-195)$$

passing through the common point of action of these forces. This is termed the *resultant* of the given system of forces.

In particular, the concurrent system of forces is *in equilibrium*, i.e., it produces no effect on the state of motion of the particle on which it acts, if

$$\mathbf{F}_1 + \mathbf{F}_2 + \dots + \mathbf{F}_N = \sum_{i=1}^N \mathbf{F}_i = 0. \quad (3-196)$$

The results (3-195) and (3-196) apply for a concurrent system of forces acting on a rigid body as well.

If the components of the forces relative to a Cartesian co-ordinate system be respectively $(F_{1x}, F_{1y}, F_{1z}), (F_{2x}, F_{2y}, F_{2z}), \dots, (F_{Nx}, F_{Ny}, F_{Nz})$, then the components of the resultant are

given by

$$F_x = \sum_{i=1}^N F_{ix}, \quad F_y = \sum_{i=1}^N F_{iy}, \quad F_z = \sum_{i=1}^N F_{iz}, \quad (3-197)$$

while the condition for equilibrium of the system reduces to the following three requirements

$$\sum_{i=1}^N F_{ix} = 0, \quad \sum_{i=1}^N F_{iy} = 0, \quad \sum_{i=1}^N F_{iz} = 0. \quad (3-198)$$

3.22.1.2 Non-concurrent forces

For a system of *non-concurrent* forces, which in the present context we will assume to be acting on a rigid body, the problem of reduction is a more involved one. This is so because in this case, the effect of the original system (made up of forces, say, $\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_N$) and the reduced system on the state of motion of the rigid body must be the same for both translational *and* angular motions of the body. We have seen that the translational motion of a system of particles is described in terms the motion of its center of mass where the rate of change of the centre of mass momentum equals the *total* force acting on the system. This means that the vector sum of the forces in the reduced system must be equal to the vector sum (say, \mathbf{F}) of the forces making up the original system.

The effect of the given system of forces on the state of angular motion of the body about any given point is represented by the total *moment* (say, \mathbf{M}) of the forces making up the system about that point. Hence, the total moment of the forces making up the reduced system about the given point must also be the same as \mathbf{M} .

In particular, for the given system to be in equilibrium, *both* \mathbf{F} and \mathbf{M} have to be zero.

Before taking up the problem of reduction of a system of forces, not necessarily concurrent, acting on a rigid body, I state below a number of useful results relating to the equilibrium of a system of concurrent forces.

3.22.2 Concurrent systems in equilibrium

3.22.2.1 Two concurrent forces in equilibrium

The condition for a pair of concurrent forces F_1 and F_2 to be in equilibrium is

$$F_1 + F_2 = 0, \text{ i.e., } F_2 = -F_1. \quad (3-199)$$

Evidently, the forces have to be equal and opposite and, being concurrent, they have the same line of action (see fig. 3-50.)



Figure 3-50: A pair of concurrent forces F_1 and F_2 in equilibrium; O is the common point of action.

3.22.2.2 Three concurrent forces in equilibrium

A. The triangle rule

For three concurrent forces to be in equilibrium, the line of action of any one of the forces has to lie in the plane containing the lines of action of the other two (reason this out). In this case, the condition (3-195) can be expressed in the following form:

The necessary and sufficient condition for three concurrent forces (say, F_1 , F_2 , and F_3) to be in equilibrium is that the directed line segments representing the corresponding vectors, placed end-to-end in order, should form a triangle, as in fig 3-51.

Here fig. 3-51(A) depicts the three forces acting at the point O. In fig. 3-51(B), the forces are represented in direction and magnitude (but *not* in their lines of action) by the directed line segments AB, BC, and CA respectively, where the tip of the third segment is seen to coincide with the initial point of the first segment, thereby completing a triangle. This satisfies the condition for the three forces to be in equilibrium. Fig. 3-51(C), on the other hand, depicts a counter-example where the line segments (AB, BC, and AC

this time) do form a triangle, but the third segment AC points in the wrong direction, thereby violating the condition of equilibrium. Indeed though the segments appear to form a triangle, they have not been placed end-to-end (initial point (the ‘tail’) of one segment coinciding with the final point (the ‘tip’) of the next) in the figure.

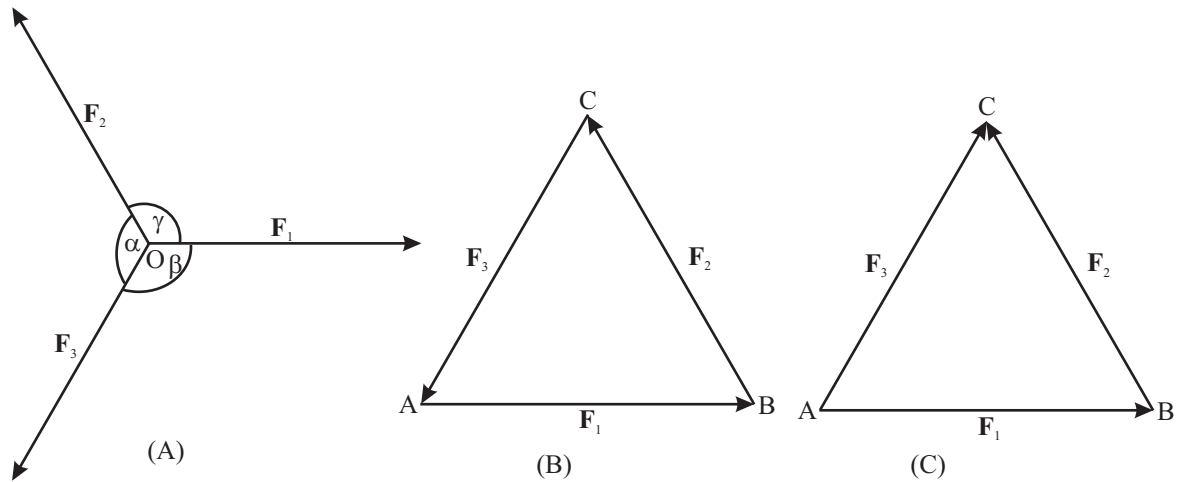


Figure 3-51: The triangle rule and the sine rule for three forces to be in equilibrium; (A) the three concurrent forces F_1 , F_2 , and F_3 ; the angles α , β , and γ are the angles featuring in the sine rule for equilibrium; (B) triangle formed by the respective directed line segments placed end-to-end in order, by parallel translation, implying that the forces are in equilibrium; (C) a counter-example, where a triangle is formed, but one of the directed line segments points the wrong way.

B. The sine rule

The necessary and sufficient condition for three concurrent forces, say, F_1 , F_2 , and F_3 , to be in equilibrium is (a) the forces have to be co-planar, i.e., their lines of action have to be contained in the same plane, and (b) the angles α , β , γ between the directed line segments representing the forces taken two at a time, as shown in fig. 3-51(A) have to satisfy

$$\frac{F_1}{\sin \alpha} = \frac{F_2}{\sin \beta} = \frac{F_3}{\sin \gamma}. \quad (3-200)$$

Here F_1 , F_2 , and F_3 stand for the magnitudes of the three forces under consideration.

Problem 3-45

Establish the triangle rule and the sine rule as stated above.

Answer to Problem 3-45

HINT: Referring to section 2.2.1.2 and to fig. 3-51(B), note that the vector extending from A to C represents, by construction, $-\mathbf{F}_3$, and also the vector $\mathbf{F}_1 + \mathbf{F}_2$, which means that the condition (3-196) is satisfied, with $N = 3$, for the three forces represented by $\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3$. In addition, recall the result that the sides of a triangle are proportional to the sines of the opposite angles, taken in order, and apply it to the three forces in figures 3-51(A), (B).

3.22.2.3 More than three concurrent forces

Think of N number of concurrent forces, say, $\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_N$ ($N > 3$). As I have mentioned above, for $N = 3$, a necessary condition for equilibrium is that the lines of action of the forces have to lie in one plane. However, for $N > 3$, this condition is no longer a necessary one - the forces may be in equilibrium even when the lines of action do not lie in one plane.

In this case, the condition (3-196) may be stated in geometrical language as follows: if the directed line segments representing the forces under consideration be placed end-to-end by parallel translation in such a way that the initial point of one segment coincides with the end point of the previous segment then these segments have to form a planar or non-planar polygon, i.e., the end point of the last segment (representing \mathbf{F}_N) has to coincide with the initial point of the first segment (representing \mathbf{F}_1). This is illustrated for a system of four forces in fig. 3-52 where, in 3-52(A), the segments form a planar polygon and in 3-52(B), they make up a non-planar polygon.

The ordering of the segments does not matter here, i.e., instead of the successive segments representing $\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4$, one can take them in the order, say, $\mathbf{F}_3, \mathbf{F}_2, \mathbf{F}_4, \mathbf{F}_1$.

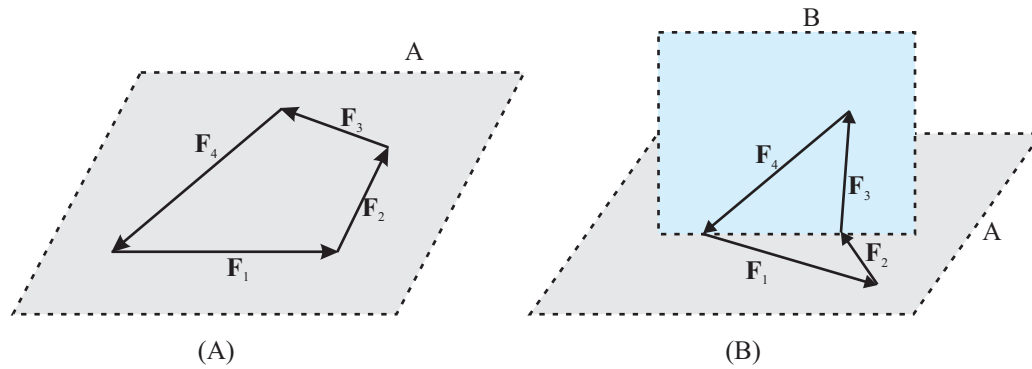


Figure 3-52: Equilibrium of four forces; in (A), the directed line segments form a planar quadrilateral, lying in the plane A, while in (B), they form a non-planar quadrilateral contained in the planes A and B.

The condition for equilibrium of a number of concurrent forces can be stated in the following equivalent way: let F_x be the sum of the resolved parts (i.e., components) of all the forces of the given system along the x-axis of a rectangular co-ordinate system, and let F_y, F_z be the corresponding sums of the resolved parts along the y- and z-axes respectively. Then the condition for equilibrium of the given system of forces is $F_x = F_y = F_z = 0$.

The following problem (see fig. 3-53) relates to an illustration of the condition for equilibrium of three concurrent forces.

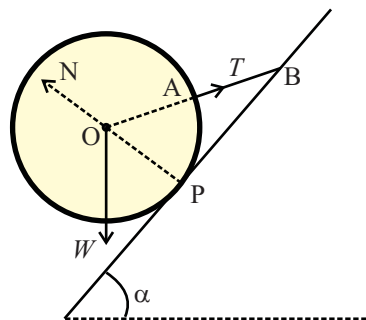


Figure 3-53: A spherical body supported on a smooth inclined plane by an inextensible string; one end of the string is attached to a point (B) on the inclined plane and the other end to a point (A) on the sphere; the weight of the sphere (W) and the reaction (N) on it of the inclined plane, both act through its center O; T denotes the tension on the string which must also pass through O; P is the instantaneous point of contact; for given values of W and α , the condition of equilibrium requires that the tension T acquires a particular value ($T = \frac{2\sqrt{3}}{3}W \sin \alpha$, if $AB=OA$).

Problem 3-46

Consider a homogeneous rigid sphere, of weight W , supported on a smooth inclined plane as in fig. 3-53 by means of a string, one end of which is attached to a point on the inclined plane and the other end to a point on the sphere. If the length of the string be equal to the radius R of the sphere, and if the plane be inclined at an angle α to the horizontal, find the tension in the string.

Answer to Problem 3-46

HINT: The tension in the string acts on the sphere along AB where A and B are the two ends of the string. The other forces acting on the spherical body are its weight W pointing vertically downward through its center, and the force of reaction (N) of the plane on it acting normally at the point of contact P (there is no tangential component to the force exerted by the plane on the sphere since the former is a smooth one), both of which also pass through the center of the sphere (O). Hence, for equilibrium, the line AB must also pass through O (reason this out). Since $AB=OA=OP=R$, it follows that $\angle OBP = \frac{\pi}{6}$. Resolving the three forces (W , N , and T , the tension) in the horizontal and vertical directions respectively, one obtains $N \sin \alpha = T \cos(\alpha - \frac{\pi}{6})$, and $N \cos \alpha + T \sin(\alpha - \frac{\pi}{6}) = W$. Eliminating N , one gets $T = \frac{2\sqrt{3}}{3} W \sin \alpha$.

3.22.2.4 Moment of concurrent forces in equilibrium

We consider a system of concurrent forces, say, F_1, F_2, \dots, F_N in equilibrium, i.e., one for which condition (3-196) is satisfied. Then the following important result holds: *for any arbitrarily chosen point O, the total moment of the forces about O is zero.*

Problem 3-47

Establish the above assertion.

Answer to Problem 3-47

HINT: Let the position vector of the point of concurrence of the lines of action of the forces relative to O be \mathbf{R} . Then the total moment about O is

$$\mathbf{M} = \mathbf{R} \times \mathbf{F}_1 + \mathbf{R} \times \mathbf{F}_2 + \dots + \mathbf{R} \times \mathbf{F}_N, \quad (3-201)$$

which is zero by (3-196).

I emphasize that the result holds regardless of the choice of the point O.

3.22.3 Reduction of a system of forces acting on a rigid body

We consider a system of forces, say, $\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_N$, where the vector sum of the forces making up the system is

$$\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2 + \dots + \mathbf{F}_N, \quad (3-202a)$$

and the total moment of the forces about any given point (say, O) is

$$\mathbf{M} = \mathbf{r}_1 \times \mathbf{F}_1 + \mathbf{r}_2 \times \mathbf{F}_2 + \dots + \mathbf{r}_N \times \mathbf{F}_N. \quad (3-202b)$$

Here $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N$ are the position vectors relative to O of points, say, P_1, P_2, \dots, P_N chosen on the lines of action of the respective forces, the locations of the points on these lines being, as we know, immaterial. The problem of reduction of this system is the one of finding a ‘simpler’ system of forces (as ‘simple’ as possible) such that the vector sum of the forces in the reduced system is \mathbf{F} and the vector sum of the moments about O of the forces in the reduced system is \mathbf{M} .

The question of simplicity of the reduced system will be clear in the context of the results to follow. In the case of a concurrent system of forces, for instance, the reduced system consists of just one single force which is evidently the simplest possible system having the same effect on the state of motion as the system one starts with.

3.22.3.1 Reduction of a pair of like parallel forces

We consider first a pair of parallel forces \mathbf{F}_1 and \mathbf{F}_2 such that the vectors representing the forces point in the same direction, as in fig. 3-54, in which A denotes the plane containing the lines of action (A_1A_2 and B_1B_2) of the two forces. Consider a third, parallel, line (C_1C_2) lying in the plane such that any transverse line (LM) is intersected by it at N

in the ratio

$$\frac{LN}{NM} = \frac{F_2}{F_1}, \quad (3-203)$$

where F_1 and F_2 denote the magnitudes of \mathbf{F}_1 and \mathbf{F}_2 respectively. Then the reduced system consists of a single force $\mathbf{F}_1 + \mathbf{F}_2$ acting along C_1C_2 (necessarily acting in the same sense as either of the two forces).

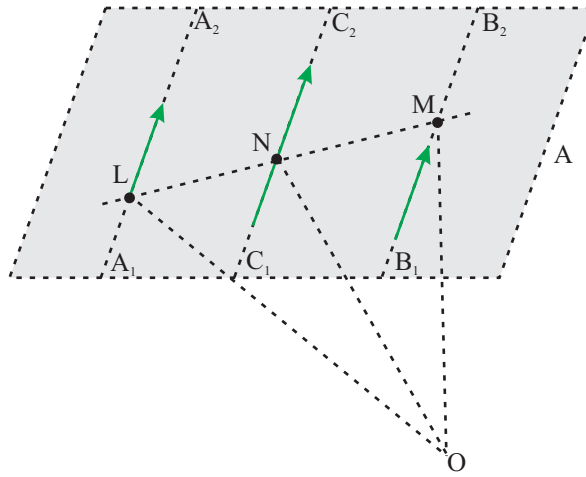


Figure 3-54: Reduction of a pair of like parallel forces; the system consisting of forces \mathbf{F}_1 and \mathbf{F}_2 along A_1A_2 and B_1B_2 respectively reduces to a single force $\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2$ along C_1C_2 parallel to the two forces; C_1C_2 intersects the transversal LM at N internally in the ratio $\frac{F_2}{F_1}$; position vectors of L and M relative to the arbitrarily chosen point O are \mathbf{r} and $\mathbf{r} + \mathbf{R}$ respectively.

It is not difficult to see why this should be so. The choice of the force \mathbf{F} already meets the requirement (3-202a). In order to see if (3-202b) is also satisfied, consider any point O with respect to which the position vectors of the points L and M (fig. 3-54) are, say, \mathbf{r} and $(\mathbf{r} + \mathbf{R})$ respectively, where \mathbf{R} stands for the vector extending from L to M . The position vector of N relative to O is then $(\mathbf{r} + \frac{F_2}{F_1 + F_2} \mathbf{R})$ (check this out), and so the moment about O of the single force $(\mathbf{F}_1 + \mathbf{F}_2)$ acting along C_1C_2 is $(\mathbf{r} + \frac{F_2}{F_1 + F_2} \mathbf{R}) \times (\mathbf{F}_1 + \mathbf{F}_2)$. This is the same as $(\mathbf{r} \times \mathbf{F}_1 + (\mathbf{r} + \mathbf{R}) \times \mathbf{F}_2)$, the total moment of the two forces about O , by virtue of the fact that \mathbf{F}_1 and \mathbf{F}_2 are parallel to each other.

Since the single force $(\mathbf{F}_1 + \mathbf{F}_2)$ along C_1C_2 to which the given pair of forces reduces is

entirely equivalent in its effects to those of the two forces \mathbf{F}_1 and \mathbf{F}_2 taken together, it is termed the *resultant* of these two forces.

3.22.3.2 Unlike and unequal parallel forces

Fig. 3-55 depicts a pair of *unlike* parallel forces of unequal magnitudes. The lines of action A_1A_2 and B_1B_2 of the forces are parallel to each other, but the forces \mathbf{F}_1 and \mathbf{F}_2 point in *opposite* directions. The line C_1C_2 divides the transversal LM *externally* at N in the ratio

$$\frac{LN}{MN} = \frac{F_2}{F_1}, \quad (3-204)$$

where, for the sake of concreteness, we assume that $F_2 > F_1$. Then the reduced system is made up of a single force $\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2$, along C_1C_2 , the magnitude of the force being $F_2 - F_1$ in the present example.

Problem 3-48

Establish the above assertion.

Answer to Problem 3-48

HINT: Make use of reasoning similar to that for a pair of like parallel forces.

3.22.3.3 Equal and unlike parallel forces : couple

Notice that, if $\mathbf{F}_2 = -\mathbf{F}_1$, i.e., the unlike parallel forces be of *equal* magnitude then the line of action of the reduced force becomes *indeterminate* (check this out). The system cannot then be reduced to a single resultant force. The vector sum of the forces being zero, the pair has no effect on the translational motion of a body, but it does have an effect on the state of angular motion since the moment about any given point, say O , is not zero. Indeed, the moment is *independent of the choice of O* .

Problem 3-49

Show that the total moment of a system of forces satisfying (3-196) about any point O is independent of the location of O.

Answer to Problem 3-49

Suppose that the moment about any point O is M. Choose a point O' with position vector, say \mathbf{R} relative to O. The moment about O' is then

$$\mathbf{M}' = (\mathbf{r}_1 - \mathbf{R}) \times \mathbf{F}_1 + (\mathbf{r}_2 - \mathbf{R}) \times \mathbf{F}_2 + \dots + (\mathbf{r}_N - \mathbf{R}) \times \mathbf{F}_N = \mathbf{M},$$

since condition (3-196) is satisfied.

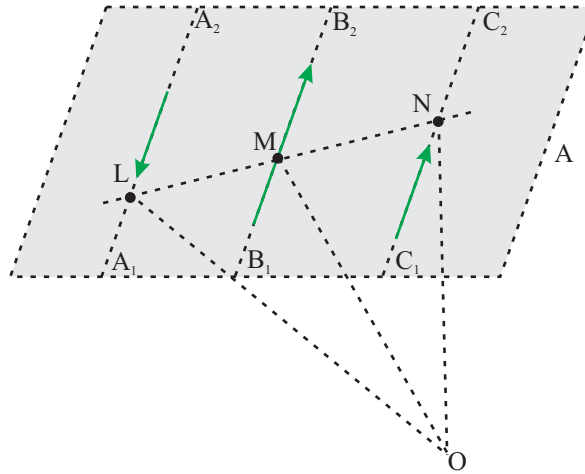


Figure 3-55: Reduction of a pair of unlike parallel forces of unequal magnitude; \mathbf{F}_1 and \mathbf{F}_2 are parallel (along A_1A_2 and B_1B_2 respectively) but point in *opposite* directions; these reduce to a single force $\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2$ along C_1C_2 parallel to the two forces, where C_1C_2 intersects the transversal LM at N *externally* in the ratio $\frac{F_2}{F_1}$ (we assume $F_2 > F_1$).

A pair of unlike parallel forces of equal magnitude, which cannot, as we have seen above, be reduced to a single resultant but which nevertheless has its effect on the state of angular motion of a rigid body on which it acts, is referred to as a *couple*.

Since the total moment about any given point of the forces constituting a couple is independent of the location of that point, it is convenient to take it on the line of action

of one of the two forces, as in fig. 3-56(A). In this figure, the lines of action of the two forces are A_1A_2 and B_1B_2 , contained in the plane A, where the forces are denoted as, say, \mathbf{G} and $-\mathbf{G}$ respectively, and O is located on the line of action (A_1A_2) of, say, \mathbf{G} . ON is the perpendicular dropped from O on the line of action of the other force (B_1B_2 in fig. 3-56(A), corresponding to $-\mathbf{G}$). Let \mathbf{R} be the position vector of N relative to O. Its magnitude ($|\mathbf{R}| = R$, the length of ON) is referred to as the *arm* of the couple. Since the moment of the force \mathbf{G} about O is zero, the total moment of the two forces constituting the couple is

$$\mathbf{M} = \mathbf{R} \times (-\mathbf{G}) = RG\hat{n}, \quad (3-205)$$

where \hat{n} denotes a unit vector normal to the plane A, as shown in the figure. Since \mathbf{M} is ultimately independent of the choice of the point O, it is termed the *moment of the couple* without reference to the location of the point O. One has to remember, though, that \hat{n} is the unit vector along $\mathbf{R} \times (-\mathbf{G})$, where \mathbf{R} is directed *from* the line of action of \mathbf{G} *to* that of $-\mathbf{G}$ in the present instance.

The effect of a couple is thus completely described by specifying the following: (a) the magnitude of the moment M , which is the product of G , the magnitude of either of the two forces constituting the couple with the arm of the couple (R), i.e., the perpendicular distance between the lines of action of the forces, (b) the unit vector \hat{n} along which the moment of the couple points.

An alternative way to describe a couple is to specify the magnitude of the moment M *along with an appropriate sign*, and any plane perpendicular to \hat{n} , referred to as the *plane of the couple* (the lines of action of the two forces are contained in one such plane). Specifying this plane is equivalent to specifying \hat{n} provided one knows *which of the two sides of the plane this unit vector points to*. This is known from the sign (+ or -) characterizing the couple: choosing any of the two sides of the plane (say, the one above the plane A in fig. 3-56) as the reference side, the sign is taken to be positive if the sense of rotation associated with the couple is anticlockwise when looked at from the reference side, and negative if this sense of rotation is clockwise (here the forces constituting the

couple are imagined to be parallel-shifted to the reference plane under consideration).

Thus, for instance, the sign is positive for the couple in fig. 3-56(A), while it is negative for 3-56(B). In both these cases the reference plane is taken to be the plane containing the lines of action of the forces. Given a choice for the reference side, the sign determines the unit vector \hat{n} : the unit vector points towards the reference side if the sign is positive, while it points in a direction opposite to the reference side if the sign is negative. In other words, the unit vector \hat{n} is related to the sense of rotation of the couple by the right hand rule (refer to sec. 2.8 for an explanation of the right hand rule).

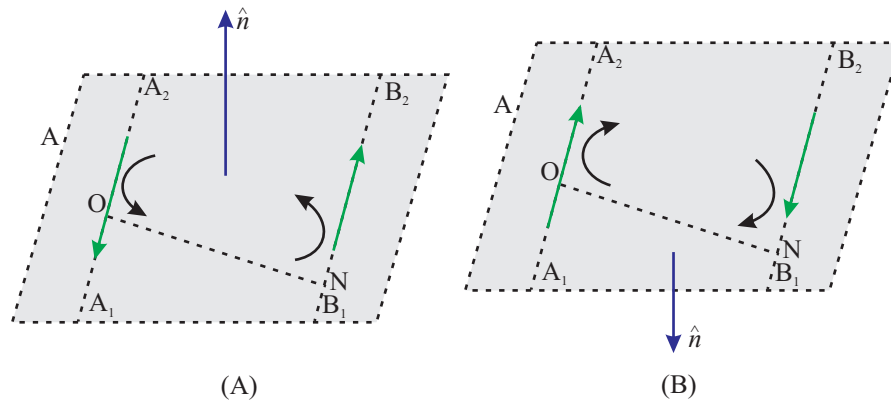


Figure 3-56: Couple made up of a pair of equal and unlike parallel forces; A is the plane containing the lines of action of the forces; two couples with opposite senses of rotation are shown in (A) and (B); in (A), O is a point located on the line of action of one of the forces, and ON is the perpendicular dropped on the other line of action; denoting the vector extending from O to N as \mathbf{R} , \hat{n} is related by the right hand rule to the sense of rotation *from* \mathbf{R} *to* the direction of the force through N; we choose the upper side of the plane A as the reference side; looked at from this side, the sense of rotation is anticlockwise in (A) and clockwise in (B); this corresponds to the sign + in (A) and - in (B).

An important corollary to follow from the above description of a couple is that one need not specify separately the magnitude (G) of either force and the arm (R) of the couple, since it is only the product ($M = RG$) that is relevant. Moreover, the plane of the couple need not be the one containing the lines of action of the two forces, since any parallel plane is as good, provided the forces are imagined to be parallel-shifted to this plane for the purpose of determining the sense of rotation, when looked at from any chosen side of the plane. Fig. 3-57 depicts two couples where the planes A and A' containing

the lines of action of the forces making up the couples are parallel to each other, and the products $M = RG$ and $M' = R'G'$ are equal. The senses of rotation for the two couples being also the same (anticlockwise when looked at from above relative to both the planes), the two couples are equivalent, and will have the same effect on the state of motion of a rigid body.

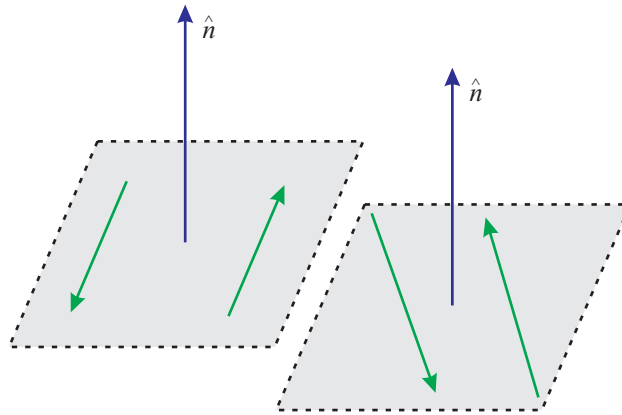


Figure 3-57: Couples with the lines of action of the forces contained in parallel planes; the sense of rotation is the same for the two couples; if, in addition, the product of the magnitude of either force in the couple and the distance between the lines of action, is also the same for the two couples, then they are equivalent in all respects; the unit vector \hat{n} depicting the direction of the moment is the same for both couples.

In summary, a couple is completely specified by the magnitude of its moment M (i.e., the product of the arm of the couple and the magnitude of either of the two constituent forces) along with the unit vector \hat{n} related to the sense of rotation associated with the couple by the right hand rule, while an equivalent description is in terms of the moment M , the plane of the couple, and its sign relative to this plane. In other words, the couple is completely specified by the vector $M\hat{n}$.

Problem 3-50

Consider a number of parallel forces $\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_N$, not necessarily acting in the same sense, where $\mathbf{F}_1 + \mathbf{F}_2 + \dots + \mathbf{F}_N \neq 0$, and where the lines of action of the forces pass through points $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N$ with reference to any chosen origin. Obtain the resultant of this system of forces.

Answer to Problem 3-50

HINT: Let \hat{n} be one of the two possible unit vectors parallel to the lines of action of the given forces, in terms of which the forces can be expressed as $\mathbf{F}_i = F_i \hat{n}$ ($i = 1, 2, \dots, N$). Here the components F_i ($i = 1, 2, \dots, N$) of the forces along \hat{n} carry their own signs, depending on the sense of \hat{n} compared to that of each of the given forces. Consider now the set of points with position vectors \mathbf{r}_i ($i = 1, 2, \dots, N$) on the lines of action of the given forces, with respect to a chosen origin. Imagine the single force $\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2 + \dots + \mathbf{F}_N = \sum_{i=1}^N F_i \hat{n}$ acting through the point $\mathbf{r} = \frac{\sum F_i \mathbf{r}_i}{\sum F_i}$ (according to the statement of the problem, the denominator differs from zero). Check that the moment of this force about the origin is the same as the vector sum of moments of the given forces. Since, moreover, the force equals the vector sum of the given forces, its moment about any other point will also be the same as the total moment of the given forces about that point (reason this out). Hence, this single force is the required resultant.

3.22.3.4 Composition of couples

Let us consider two couples of moments M and M' , such that the plane containing the forces making up the former is parallel to that for the latter.

Since the plane of the couple may mean any plane parallel to the one containing the forces making up the couple, the couples M and M' being considered here may be said to belong to the same plane.

Here the moments are assumed to carry their appropriate signs. The two couples in this case may be reduced to a single couple in the same plane as either of the two, with a moment $M + M'$. Depending on the signs of M and M' , this may mean any one of the four quantities $\pm|M| \pm |M'|$.

If, on the other hand, the planes of the two couples are inclined to each other, then they reduce to a couple in yet another plane. In this case it is more convenient to represent the couples as vectors \mathbf{M} and \mathbf{M}' , when the reduced couple is given by the vector $\mathbf{M} + \mathbf{M}'$.

In reality, as explained in section 2.6.1, the moment of a couple is a *pseudo*-vector, or an axial vector, rather than a (polar) vector.

Thus, two couples can always be reduced by composition to a single couple. Evidently, one can generalize this by saying that any number of couples can be reduced to a single couple.

3.22.3.5 A couple and a force in a parallel plane

Fig. 3-58 depicts a force F and a couple M , where the line of action (A_1A_2) of the force is parallel to the plane (say, A) containing the two forces the couple is made up of. In this case the force and the couple may be said to be *co-planar*. Can this system be reduced to a simpler one?

Consider a plane B parallel to A containing the line of action of the force F , and a force represented by the same vector F as the one along A_1A_2 , but now along a different line, say, B_1B_2 in the plane B , such that the moment of the force about any point (say, O) on A_1A_2 is M . The distance between the lines A_1A_2 and B_1B_2 is given by $d = \frac{|M|}{F}$. The line B_1B_2 is to be chosen on that side of A_1A_2 for which its moment about O is in the same sense as the sense of rotation corresponding to the given couple.

The given system of forces consisting of the force F along A_1A_2 and those constituting the couple M then reduces to the single force F along the line B_1B_2 , which is termed the *resultant* of the given system.

Problem 3-51

Check the above statement out.

Answer to Problem 3-51

By construction, the moment of the resultant about O is the same as the total moment of the given system of forces. In addition, the vector representing the resultant force is the vector sum

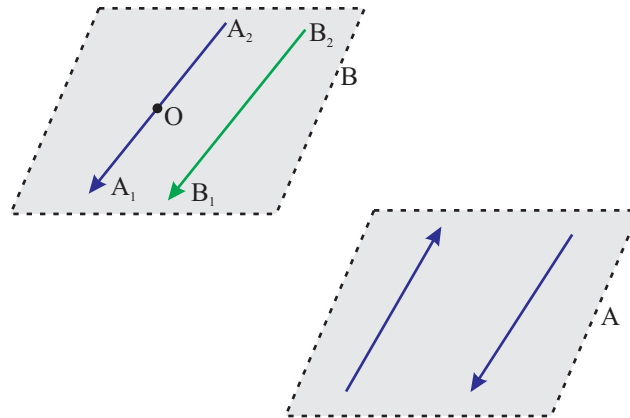


Figure 3-58: A system made up of a force F along A_1A_2 and a couple M in a parallel plane A ; the plane B is drawn parallel to A ; the system reduces to a single force F along B_1B_2 in the plane A ; B_1B_2 lies on that side of A_1A_2 for which the sense of rotation of the resultant force about any point O on A_1A_2 is the same as the sense of rotation of the couple.

of those in the given system.

The above statement admits of a converse in the following sense.

Let F be any given force acting along the line A_1A_2 , and O any given point. Then this force can be replaced with (a) a force represented by the same vector F but with its line of action passing through O , and (b) a couple in a plane containing the line A_1A_2 and the point O , the moment of the couple being the same as the moment of the force F about the point O (i.e., $M = r \times F$, where r stands for the position vector of any point on A_1A_2 relative to O). Looked at from any chosen side of the said plane, the sense of rotation of the couple will be the same as that of the force F with respect to O .

3.22.3.6 Reduction of a system of co-planar forces

Let us start from two co-planar forces, say, F_1 and F_2 . If they form a pair of unlike parallel forces of equal magnitudes, they form a couple, say M , the plane of the couple being the same as that of the given forces. Else, they reduce to a single resultant force, say, F , again in the same plane (refer to the cases of concurrent forces, like parallel forces, and unlike parallel forces of unequal magnitude). Now consider a third co-planar force, say F_3 . In the case of the forces F_1 and F_2 forming a couple, this couple

(M), together with the force F_3 , reduce to a single force in the plane of the given forces.

On the other hand, if F_1 and F_2 reduce to a single resultant then this resultant (F), together with the force F_3 will again reduce to either a single resultant force or a couple, once again in the plane of the given forces. In any case, then, a system of three co-planar forces reduces to either a single resultant force or to a couple in the same plane. One can now bring in a fourth force and, by the same reasoning, obtain either a single resultant force or a couple. Evidently, this process of reasoning can be repeated for any finite number of co-planar forces. We thereby arrive at the following important conclusion:

A system of co-planar forces reduces either to a single resultant force or to a couple, in the same plane as the given system of forces.

The system of forces for which the above result holds can be generalized to a certain extent, namely a set of forces lying in a single plane, say, A, together with a set of couples in the same plane. The pair of forces constituting any one of this set of couples need not lie in the plane A, it being only necessary that they lie in a plane parallel to A. All the couples can then be said to be *in the plane A*.

3.22.3.7 Reduction of non-coplanar forces: wrench

Finally, we consider the problem of reduction of a system of *non-coplanar* forces acting on a rigid body. Here one arrives at the following result:

A system of forces, not necessarily co-planar, can be reduced to a couple and a force perpendicular to the plane of the couple.

A system of forces made up of a couple and a force perpendicular to the plane of the couple, is referred to as a *wrench*. Figure 3-59 depicts a wrench made up of a couple of moment, say, M in the plane A, and a force, say F perpendicular to A. If the given system of forces be concurrent then M reduces to zero. On the other hand, if the system

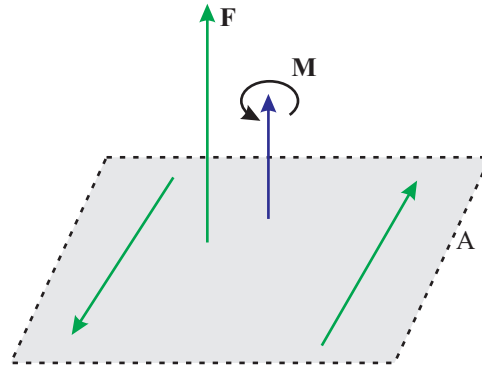


Figure 3-59: A wrench, made up of a couple M in the plane A , together with a force F perpendicular to A .

of forces be co-planar, then either M or F reduces to zero.

The following problem relates to a pair of forces whose lines of action are skew to each other (fig. 3-60) and demonstrates its equivalence to a wrench.

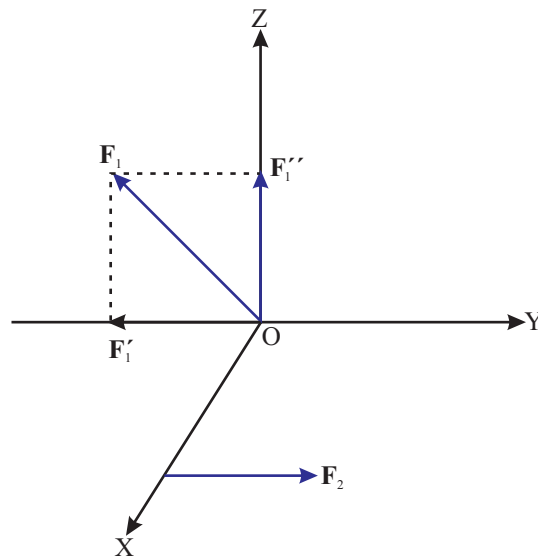


Figure 3-60: Two forces $\mathbf{F}_1 = -A\hat{j} + A\hat{k}$ and $\mathbf{F}_2 = A\hat{j}$ with their lines of action skew to each other; replacing \mathbf{F}_1 with the two forces $\mathbf{F}'_1, \mathbf{F}''_1$ the system is seen to reduce to a wrench, made up of the force \mathbf{F}''_1 along with a couple G in the x - y plane.

Problem 3-52

Consider the forces \mathbf{F}_1 and \mathbf{F}_2 acting on a rigid body where $\mathbf{F}_1 = -A\hat{j} + A\hat{k}$, whose line of action passes through the origin of a rectangular co-ordinate system and $\mathbf{F}_2 = A\hat{j}$ passing through the point with co-ordinates $(1, 0, 0)$ (see fig. 3-60), A being a given constant. Show that the pair of forces is equivalent to a wrench made up of a force $A\hat{k}$ passing through the origin, and a couple in the x-y plane. Find the moment of the couple. What is the moment of the pair of forces about the point $(0, 1, 0)$?

Answer to Problem 3-52

HINT: The given system is equivalent to the three forces $\mathbf{F}'_1 = -A\hat{j}$ and $\mathbf{F}''_1 = A\hat{k}$, both acting through the origin, and $\mathbf{F}_2 = A\hat{j}$ acting through $(1, 0, 0)$. Of these, \mathbf{F}'_1 and \mathbf{F}_2 form a couple $\mathbf{G} = A\hat{k}$ in the x-y plane. Thus the given pair of forces reduces to a wrench consisting of the force $\mathbf{F}''_1 = A\hat{k}$ acting through the origin, and the couple \mathbf{G} in the x-y plane. The moment of the force \mathbf{F}''_1 acting through the origin about the point $(0, 1, 0)$ is $(-\hat{j}) \times A\hat{k} = -A\hat{i}$. Adding to this the moment of the couple $\mathbf{G} = A\hat{k}$, the moment of the given system about the point $(0, 1, 0)$ is seen to be $A(-\hat{i} + \hat{k})$. The magnitude of the moment is $\sqrt{2}A$.

3.23 Static and dynamic friction

3.23.1 Introduction

Fig. 3-61 depicts a block B at rest on a horizontal plane S (say, flat ground), where a small force F is applied to the block in the horizontal direction as shown in the figure. Experience tells us that, if the applied force F is sufficiently small, the block does not move on the plane S, which implies that an equal and opposite force must be acting on the block by virtue of its contact on the ground. As the applied force F is made to increase so as to cross a certain limiting value (say, F_0), the block starts moving in the direction of the force. In other words, the opposing force brought into play due to the contact of the block with the ground cannot exceed the limiting value F_0 , where the latter is found to depend in a complex manner on the nature of the two surfaces in contact, as also on the force of normal reaction (N) applied by the ground on the block.

While the opposing force referred to above acts when the block is at rest on the ground,

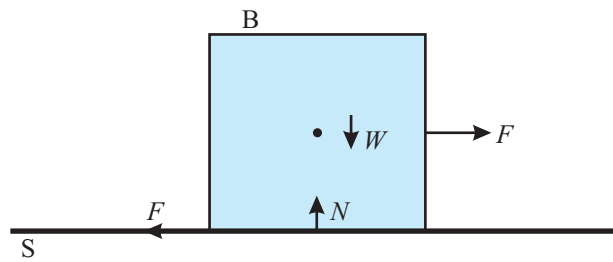


Figure 3-61: Illustrating the force of friction between two bodies; a force F applied to the block B resting on the horizontal surface S causes an opposing force, due to friction, to appear by virtue of its contact with the surface; the opposing force balances the applied force up to a limiting value of F ; if F exceeds this limiting value, the block starts moving; while the applied force F and the frictional force are both horizontal in the present instance, the weight W of the body and the reaction force N of the ground act in the vertical direction, with the latter two balancing each other.

a similar resistive force is found to be brought into play when the block is made to move on the ground as well, where this resistive force acts in a direction opposite to that of the velocity of B relative to S. Thus, if the block moves on the ground along the direction shown by the arrow in fig. 3-62, then a certain force (say, F' , not marked in the figure) acts on it by virtue of its contact of the ground where this force is seen to possess the following characteristic features: (a) it depends in a complex way on the nature of the two surfaces in contact, (b) it depends, to a small extent, on the relative velocity between the two surfaces (the velocity of the block on the ground in the present instance), and (c) it depends on the degree to which the block presses against the ground, i.e., on the magnitude of the forces of action and reaction between the two surfaces in contact, where these forces act in a direction normal to the common tangent plane to the two surfaces.

This phenomenon of a resistive force coming into play between two surfaces in contact is referred to as *friction*, where one commonly distinguishes between *static* and *dynamic* friction, corresponding to the situations depicted in figures 3-61 (with the block B resting on the surface S), and 3-62 (with B in motion relative to S) respectively.

The force of dynamic friction turns out to be a variable one in the early instants after the relative motion is initiated, showing rapid oscillations till a steady motion sets in.

The study of friction can be broadly divided into two parts. Of these, one relates to the formulation of the characteristic features of frictional forces in a phenomenological way (i.e., one based on empirical observations) and making use of these in solving for the motions of bodies observed in daily experience and in engineering practice. The other part relates to looking for the *origin* of frictional forces in more fundamental terms and explaining the observed features of these forces.

The features of friction I have briefly referred to above and will consider in greater details in sections 3.23.2 and 3.23.3 all relate to what is referred to as *dry* friction where there is no layer of any liquid or gas in between the two surfaces in contact. Thin films of liquid or vapor between the surfaces are, however, unavoidable in practice and their presence, while causing some deviations from the rules pertaining to ideally dry friction, will not be included from our considerations relating to dry friction. *Wet friction* will be considered briefly in sec. 3.23.6

In the following, we consider a body B in contact with the surface S of some other body, where the force of friction between the two bodies acts along the surface of contact between B and S.

3.23.2 Static friction

A basic feature relating to static friction, mentioned in sec. 3.23.1 is that, for a given pair of surfaces in contact, the force of friction opposing the applied force tending to cause a relative motion of the two surfaces, cannot exceed a certain limiting value depending on the magnitude (say, N) of the force of normal reaction between the two, where the latter denotes the magnitude of the forces of action and reaction between the two bodies in contact, acting in a direction normal to their common tangent plane. This *force of limiting friction* (say, F_0) is related to N as

$$F_0 = \mu_s N, \tag{3-206}$$

where μ_s is a constant for the given pair of surfaces in contact, and is termed the *coefficient of static friction*.

The force of static friction is brought into play only if a force is applied on B (we assume for the sake of simplicity that the surface S on which B moves (or tends to move) is a fixed one, being the boundary surface of some other heavy body), and acts in a direction opposite to this force, tending to resist the relative motion between the two bodies in contact. Along with the force of friction acting on the body B under consideration, there will also be an equal and opposite force acting on the other body (with which B is in contact).

The coefficient μ_s depends on the nature of the two surfaces in contact. For instance, it is, in general, found to decrease as the degree of roughness of either of the two surfaces is made to decrease by polishing or smoothening (however, too much of smoothening may result in an increase in friction). It also depends sensitively on the presence of contaminants (like, say, a drop of oil, or a film of oxide coating) on the surfaces.

The relation (3-206) expressing the proportionality between the force of limiting friction F_0 and the normal reaction N is referred to as *Amontons' law* in friction.

A fact of basic importance in connection with Amontons' law is that *the force of limiting friction (F_0) does not depend on the area of contact* between the two bodies under consideration. To be more precise, the frictional force is independent of the *apparent* area of contact, i.e., the area over which the two bodies are seen to overlap. In contrast, we will consider below the *actual* area of contact which is smaller than the apparent area, and on which the frictional force of limiting friction (F_0) does depend, since the actual area of contact increases in direct proportion to N . The coefficient μ_s , however, is independent of both the apparent and actual areas of contact.

3.23.3 Dynamic friction

The basic rule of dynamic friction, which comes into play when two bodies in contact with a common surface (S in the present instance) are in relative motion, can again be

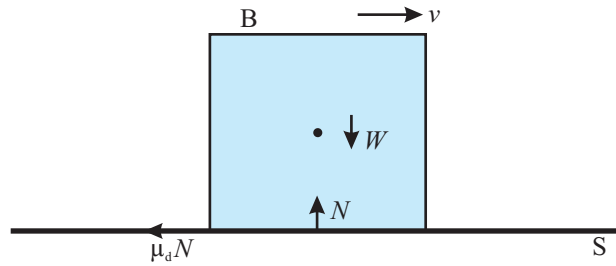


Figure 3-62: Illustrating dynamic friction on a block B moving on a surface S with a relative velocity v ; the frictional force opposing the motion is $\mu_d N$, where N is the force of normal reaction exerted by S on B, and μ_d is the coefficient of dynamic friction; W denotes the weight of the block.

expressed in a form similar to eq. (3-206), namely

$$F = \mu_d N, \quad (3-207)$$

where, in general, the coefficient of dynamic friction (μ_d), differs from μ_s , being a function of the relative velocity v between the surfaces. As already mentioned, the force of dynamic friction acts in a direction opposite to that of the relative velocity between the bodies in contact.

In general, the dependence of μ_d on v looks as in fig. 3-63(A) where, for a certain range of the relative velocity above zero, the coefficient μ_d is almost independent of v . This is why the coefficient μ_d is, at times, said to be independent of v , this velocity-independence of the force of dynamic friction being referred to as *Coulomb's law* in friction.

For higher values of the relative velocity, however, the force of friction (and the coefficient μ_d) is first found to decrease with v and then to increase, as shown in the figure. However, for the sake of simplicity, the variation of μ_d with v can be represented graphically as in fig. 3-63(B), in ideal accordance with Coulomb's law.

As in the case of static friction, the coefficient μ_d depends in a complex way on the nature of the two surfaces in contact (which includes their degree of roughness and the presence of contaminants), and is independent of the apparent or the actual area of contact between the two bodies under consideration. The force of friction depends on the

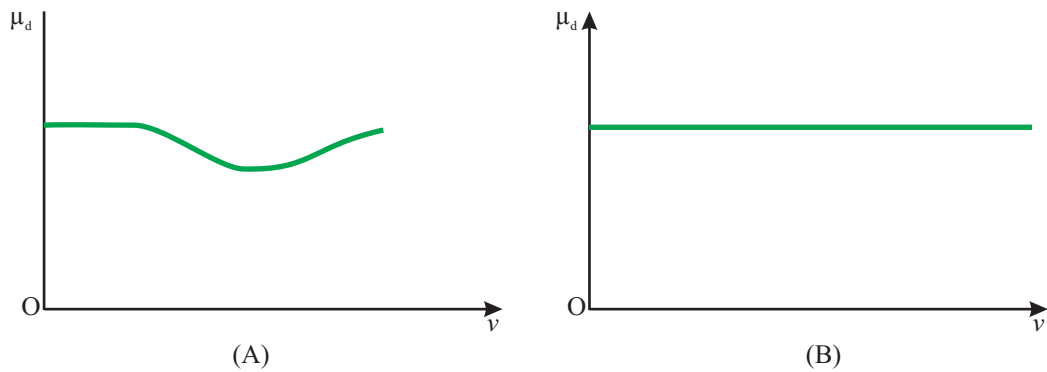


Figure 3-63: The dependence of the coefficient of dynamic friction (μ_d) on the relative velocity v ; (A) actual, where μ_d at first remains constant, then decreases slightly, finally increasing with v ; (B) ideal, where μ_d remains constant.

actual area of contact which, in contrast to the apparent area, increases in proportion to the normal reaction N , which explains the presence of the factor N in eq. (3-207).

The frictional force between any two bodies in contact is brought into play due to the mutual interaction between the atoms and molecules of the two bodies near their area of overlap. Consequently, the force of friction on any one of the two bodies brings in an equal and opposite frictional force acting on the other body as well.

At the cost of repetition, it is to be emphasized that the ‘laws’ of static and dynamic friction are of an empirical nature, based on regularities observed in real-life situations including those in engineering practice, as also in laboratory experiments. These laws are only partly explained in terms of more fundamental principles relating to molecular interactions between bodies.

Historically, the work of Amontons relating to the laws of friction precedes that of Coulomb, who based his researches on earlier findings of Amontons, systematizing and extending the principles arrived at by the latter. Thus, there is a considerable overlap between results enunciated by the two pioneers and the practice of associating the name of one or the other of them with this or that ‘law’ of friction is, to some extent, arbitrary. Besides, one cannot rule out the possibility that the same ‘laws’ were arrived at by other workers not so well-recognized in history.

The frictional force brought into play when two bodies slide on one another, makes necessary the expenditure of energy in the form of *work* to bring about the relative displacement between the two. Considering a small displacement δx of either of the two bodies over the other, the work to be performed *against* the frictional force is given by

$$\delta W = -F\delta x = \mu_d N\delta x, \quad (3-208)$$

which is always positive since the frictional force on either of the two bodies acts in a direction opposite to its displacement, i.e., F and δx carry opposite signs (this means that a negative sign is understood in eq. (3-207)).

The energy thus expended in the form of work eventually gets dissipated in the materials of the two bodies as heat (see sec. 3.23.5) or, stated more precisely, in the form of *thermodynamic internal energy* (see chapter 8).

When a body is made to slide on ice, the heat generated results in the melting of ice into water locally at the area of contact, which acts as a *lubricant*, facilitating the sliding process.

The lowering of the melting point due to the pressure exerted by the sliding body on ice (see sec. 8.22.3) is, in general, responsible for the facilitated sliding to a lesser extent as compared to the effect of the melting due to the heat generated by friction.

The following problem illustrates the application of the rules of static and dynamic friction (under idealized conditions) in the case of two blocks placed on the ground, one above the other, where a force F is applied on the upper block as shown in fig. 3-64.

Problem 3-53

Two blocks A and B, whose weights are respectively W_1 and W_2 , are placed on the ground, with A resting on B. The coefficient of friction between A and B is μ_1 while that between B and the ground is μ_2 , where it is assumed that, in either case, the coefficients of static and dynamic friction are

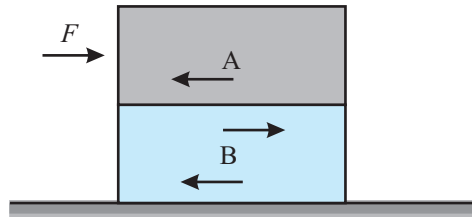


Figure 3-64: Blocks A and B resting on the ground, one above the other; a force F is applied to the upper block A; forces of friction appear on the blocks as shown by arrows; depending on the magnitude of F and the coefficients of friction μ_1 and μ_2 between A and B and between B and the ground, various possibilities arise regarding the motion of the blocks.

the same. If a force F is applied to the upper block, find the conditions under which (a) both the blocks continue to rest on the ground, (b) the upper block A moves with an acceleration, say, f_1 , over the lower block B, which remains static, (c) both the blocks move together with an acceleration, say, f_2 , and (d) both A and B move forward, with the acceleration (say, f_3 ,) of the former being greater than that (f_4) of the latter. Find the values of f_1 , f_2 , f_3 , f_4 in the respective cases.

Answer to Problem 3-53

SOLUTION: (a) Since there is no relative motion between the blocks, the frictional force between A and B must be less than $\mu_1 W_1$. Since this force, acting on A, balances F , one must have $F < \mu_1 W_1$. This force also acts on B in the forward direction (i.e., in the direction of F) and is balanced by the frictional force exerted by the ground, which is less than $\mu_2(W_1 + W_2)$. Thus, $F < \mu_2(W_1 + W_2)$. In other words, the required condition is $F < \min(\mu_1 W_1, \mu_2(W_1 + W_2))$.

(b) Since A moves over B, $F > \mu_1 W_1$, where $\mu_1 W_1$ is the force of friction between A and B, acting on A in the backward direction (opposite to F), and on B in the forward direction. Since B is stationary on the ground, $\mu_1 W_1 < \mu_2(W_1 + W_2)$. Thus, the required condition is $\mu_1 W_1 < \min(F, \mu_2(W_1 + W_2))$. The acceleration of A is $f_1 = \frac{F - \mu_1 W_1}{W_1}g$, where g stands for the acceleration due to gravity.

(c) Let the force of friction between A and B be F' , which acts on A in the backward direction and on B in the forward direction. Since there is no relative motion, $F' < \mu_1 W_1$. Again, considering the two blocks together, since they jointly move on the ground, $F > \mu_2(W_1 + W_2)$. Since the blocks have the same acceleration, $\frac{F - F'}{W_1} = \frac{F' - \mu_2(W_1 + W_2)}{W_2}$, i.e., $F' = \frac{FW_2}{W_1 + W_2} + \mu_2 W_1$. Collecting all these, the required condition turns out to be $\mu_2(W_1 + W_2) < F < (\mu_1 - \mu_2)\frac{W_1}{W_2}(W_1 + W_2)$ (this requires $\mu_2(W_1 + W_2) < \mu_1 W_1$). The common acceleration is $f_2 = \frac{F - F'}{W_1}g$.

(d) Since there is relative motion between A and B, the frictional force between them is $\mu_1 W_1$, acting in the backward direction on A and forward direction on B. The frictional force on B exerted by the ground is $\mu_2(W_1 + W_2)$ in the backward direction. Since A moves forward relative to B, one has $\frac{F - \mu_1 W_1}{W_1} > \frac{\mu_1 W_1 - \mu_2(W_1 + W_2)}{W_2}$. This gives $F > (\mu_1 - \mu_2) \frac{W_1}{W_2} (W_1 + W_2)$. The required condition is thus $F > \max((\mu_1 - \mu_2) \frac{W_1}{W_2} (W_1 + W_2), \mu_1 W_1)$, $\min((\mu_1 - \mu_2) \frac{W_1}{W_2} (W_1 + W_2), \mu_1 W_1) > \mu_2(W_1 + W_2)$. The acceleration of A is $f_3 = \frac{F - \mu_1 W_1}{W_1} g$, and that of B is $f_4 = \frac{\mu_1 W_1 - \mu_2(W_1 + W_2)}{W_2} g$.

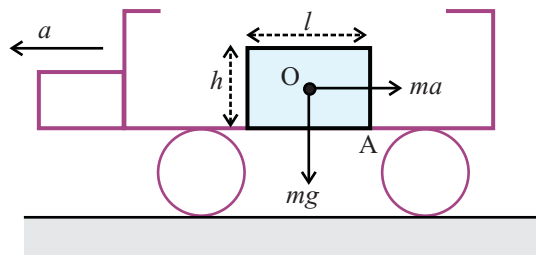


Figure 3-65: A rectangular block of length l and height h resting on the floorboard of an accelerating vehicle; if the acceleration a is sufficiently large, the block topples about its rear edge A, where it is assumed that the coefficient of friction is sufficiently large to prevent sliding; as the block topples, the reaction on it exerted by the floor of the vehicle acts through A; in a frame of reference fixed on the vehicle, the other forces on the block are its weight mg acting vertically downward through the center O and the inertial force ma acting horizontally through O in a direction opposite to that of the acceleration.

Problem 3-54

A homogeneous rectangular block rests on the floorboard of a vehicle moving with an acceleration a . If the length of the block along the direction of motion of the vehicle be l (see fig. 3-65) and its height be h , find the acceleration above which the block topples about its rear edge. Assume the coefficient of friction between the block and the floorboard to be sufficiently large.

Answer to Problem 3-54

Consider a frame of reference fixed on the accelerating vehicle, in which a pseudo-force ma acts on the block in the backward direction through its center of mass O (see sections 3.10.3, 3.21), while its weight mg acts vertically downward through O. Assuming the coefficient of friction between the block and the floor of the vehicle to be sufficiently large, sliding of the former on the latter is prevented while the block tends to topple about its rear edge A (fig. 3-65). As the block topples,

the force of reaction exerted by the floor of the vehicle acts through A. Taking moments about A, equilibrium is just possible if $mg\frac{l}{2} = ma\frac{h}{2}$ (which follows from the basic principle that, for a system of forces to be in equilibrium, the vector sum of the moments of the forces about any arbitrarily chosen point has to be zero, and hence each component of the vector sum is also to be zero - reason this out). The block will topple if $a > g\frac{l}{h}$.

Problem 3-55

Sand grains are dropped vertically at a uniform rate m onto a horizontal conveyer belt, to be carried away by the latter, which is made to move horizontally with uniform speed u with the help of a motor (fig. 3-66). Work out the rate at which the kinetic energy of the sand collected on the belt increases, and the rate at which work is being done to move the belt along with the sand being dropped on it. Account for the difference between the two.

Answer to Problem 3-55

SOLUTION: If M be the mass of sand on the belt at time t , then rate of collection of sand on the belt is $m = \frac{dM}{dt}$. In a small time interval δt , $m\delta t$ mass of sand develops a horizontal velocity u , which means that the rate of increase of kinetic energy is $\frac{1}{2}mu^2$. The increase in momentum in the horizontal direction in time δt is $mu\delta t$. Hence the rate of change of momentum, which gives the force required to drive the belt, is mu . Since in time δt the belt moves a distance, $u\delta t$ along with its load, the work done is $mu^2\delta t$. Thus the required driving power is mu^2 . Only half of this power is expended to increase the kinetic energy of sand collecting on the belt.

The remaining expenditure of energy is accounted for by friction. Assuming that the sand falling on the belt does not slip backward, as implied in the problem statement, and that the frictional force per unit mass is f , the time required by sand to accelerate from 0 to u is $\frac{u}{f}$. At any instant of time there is some mass of sand with velocity between 0 and u , undergoing acceleration. Considering sand grains with velocity between v and $v + \delta v$, they started at an earlier time in an interval between $\frac{v}{f}$ and $\frac{v+\delta v}{f}$, i.e., the mass of sand in this velocity range is $m\frac{\delta v}{f}$, the frictional force on this mass being $m\delta v$. Since the relative velocity is $u - v$, the rate of work done against friction is $m(u - v)\delta v$, implying that the total rate of work done against friction is $m \int_0^u (u - v)dv = \frac{1}{2}mu^2$.

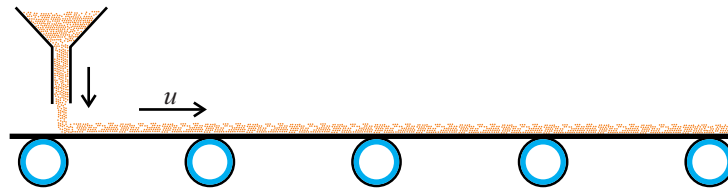


Figure 3-66: Sand grains are dropped vertically at a uniform rate onto a horizontal conveyer belt, to be carried away by the latter; the belt, powered by a motor, is made to move horizontally with uniform speed u ; half of the the power required to drive the belt along with the accumulating sand is expended for imparting kinetic energy to the sand grains while the other half is expended to overcome the frictional drag exerted on the belt the sand.

3.23.4 Indeterminate problems in statics: the ladder problem

The principles of statics leading to the conditions of equilibrium of a system of forces acting on a body brought into play by the action of other bodies on it, as enunciated in sec. 3.22, are applicable in diverse real-life situations including those in engineering practice.

However, there exist a class of problems where the forces acting on one or more bodies at rest cannot be determined from the conditions of equilibrium alone since the number of unknown force components to be determined is found to exceed the number of equations following from the conditions of static equilibrium. These are referred to as *statically indeterminate* problems. In a typical such indeterminate problem, the complete determination of all the relevant forces acting on a body requires that other principles be invoked, relating to the structural properties of the bodies exerting forces on one another. In other words, the conditions of equilibrium of a system of forces is to be complemented with equations determining the forces of deformation associated with elastic strain and stress in one or more bodies (refer to chapter 6 for an introduction to elastic properties of deformable bodies) since idealized assumptions relating to perfectly rigid bodies prove insufficient in determining the forces in question.

Situations of such type are found to arise in structural engineering where one needs to determine the conditions of equilibrium of structures made up of bodies in contact. The contacts may be of various types (e.g., clamped and pinned joints), and stresses are developed in the bodies, depending on the constraints imposed by these. Typically, one

needs numerical analysis for the determination of the forces since the relevant equations to solve are numerous and complex.

A well-known indeterminate problem frequently met with in real life situations is that of the inclined ladder (see below) which is, at the same time, important from a social perspective since workmen climbing up ladders inclined against walls meet with accidents caused by erroneous settings that cause the ladders to slide and fall. More generally, the reaction forces acting on a body in multiple contact with other bodies are often indeterminate when looked upon as problems of static equilibrium of forces between rigid bodies without reference to their structural properties. As a simple example, consider a heavy and uniform rigid block resting on two rigid supports as shown in fig. 3-67, and imagine that a horizontal pull is applied at the point P by means of a force F . Assuming the contacts on the supports to be at points A, B, one can obtain the normal reactions at A, B as $N_1 = N_2 = \frac{W}{2}$ (reason this out), while the frictional forces F_1 , F_2 at A and B are required to satisfy the relation $F_1 + F_2 = F$ (along with $F_1 \leq \mu_1 N_1$, $F_2 \leq \mu_2 N_2$) as long as the system is in equilibrium. This constitutes an indeterminate problem. The forces F_1, F_2 get determined only in the case of *limiting equilibrium*, i.e., the one when $F = \mu_1 N_1 + \mu_2 N_2$. In order to determine all the forces developing at the points A, B, one needs to have more information regarding how the contacts are set up and (possibly) regarding the elastic properties of the material of the block as well (assuming that the bodies providing the supports are rigid ones).

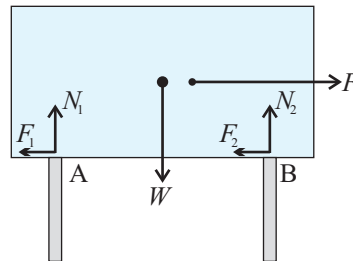


Figure 3-67: The example of an indeterminate problem in statics; a uniform rigid body of weight W is supported rigidly at A and B, and a horizontal pull is applied on it at P; the forces of reaction F_1 and F_2 at A, B remain indeterminate as long as the system is in equilibrium, and get determined only when the equilibrium is a limiting one; a complete determination of all the reaction forces at A, B requires further information on how the supports are set up and (possibly) on the elastic properties of the material of the body as well.

3.23.4.1 The ladder problem

Fig. 3-68 depicts a ladder, assumed to be a rigid body, with its lower end on a rough floor at A and upper end resting against a rough wall at B. The weight W of the ladder acts vertically downwards at its mid-point while another weight W' rests on the ladder at some point C such that $PR = \alpha l$, where $l = PQ$ is the length of the ladder (W' may represent a workman climbing up from A to B, C being his instantaneous position). Assuming that the entire system is in equilibrium, one obtains the following relations involving the normal and tangential forces of reaction at A, B (refer to the figure)

$$\begin{aligned} N_2 = F_1, \quad N_1 + F_2 = W + W', \quad N_2 l \sin \theta + F_2 l \cos \theta &= \frac{1}{2} W l \cos \theta + \alpha W' l \cos \theta \\ F_1 \leq \mu_1 N_1, \quad F_2 \leq \mu_2 N_2, \end{aligned} \quad (3-209)$$

(check this out) where θ stands for the inclination of the ladder to the horizontal and μ_1, μ_2 denote the coefficients of friction at the floor and the wall respectively. This is clearly an indeterminate system since, with given values of the various parameters, there are four unknown forces and only three equations representing the conditions of equilibrium.

There exists a large body of literature addressing the ladder problem, all looking at ways for making the problem determinate.

I recommend the following two papers for a good overview of the problem:

1. M.P. Silverman, 'Reaction Forces on a Fixed Ladder in Static Equilibrium: Analysis and Definitive Experimental Test of the Ladder Problem', World Journal of Mechanics, vol.8, no.9, 311-342 (1918).
2. M.P. Silverman, 'The Role of Friction in the Static Equilibrium of a Fixed Ladder: Theoretical Analysis and Experimental Test', World Journal of Mechanics, vol.8, no.12, 445-463 (1918).

A common assumption made in elementary text-books is $\mu_2 = 0$ which, however, is an arbitrary one. In order to determine all the four reaction forces at A and B, one has

to pay attention to contingent factors such as what type of strains are developed in the ladder and what actually happens at the two contacts. Elaborate experiments have been conducted, determining the actual reaction forces developed, and several models have been set up where several specific types of strain are assumed to characterize the deformation of the ladder. What is more, *turning moments* developing at the contacts may also have to be taken into account to fully describe the equilibrium of the ladder.

However, a considerable number of experimental tests indicate that the equilibrium of the ladder is consistent with a relatively simple additional constraint under which all the reaction forces get determined:

$$F_2 = \mu_2 N_2. \quad (3-210)$$

In other words, the equilibrium of the ladder is achieved when the upper contact is in a position of limiting friction while the lower contact is still away from limiting friction. Assuming that the ladder is placed by first bringing its lower end in contact with the floor and then making the upper end touch the wall, what actually happens (as found in a majority of experiments) is that, for a given position of the climber (who may be anywhere between A and B), the upper end slips on the wall till the condition of limiting friction (eq. (3-210)) is achieved. At the same time, the ladder develops a compressive longitudinal strain (refer to chapter 6 for basic ideas in elasticity) and is also likely to get bent, with the lower end away from limiting friction ($F_1 \leq \mu_1 N_1$). As the climber moves towards the top of the ladder, α keeps on increasing, and the ladder continues to remain in equilibrium for each position of the climber (we assume that the climber moves up infinitesimally slowly) till a value of α is reached when the lower contact reaches the state of limiting friction ($F_1 = \mu_1 N_1$) for a given value of θ . This gives the condition of limiting equilibrium for the ladder as a whole. In the range of situations in which this description is valid, the elastic properties of the ladder do not directly enter into the equations determining the reaction forces (up to a good degree of approximation).

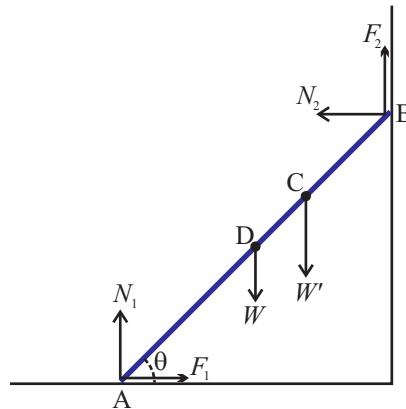


Figure 3-68: Illustrating the ladder problem, an indeterminate problem in statics; a ladder of length l and weight W is brought into contact with the ground at A, whereafter its other end B is made to rest against a vertical wall; a load W' (a climber, for instance) is assumed to be located at C, where $AC=\alpha l$; the weight W acts through the middle point D; reaction forces with components N_1, F_1 and N_2, F_2 are produced at A and B respectively, but all these components cannot be determined from the conditions of static equilibrium alone (equations (3-209)), unless additional information is made use of; one such piece of information, inferred from experimental observations, is that the top of the ladder slips down till the condition of limiting equilibrium is satisfied at B, while A remains fixed without reaching the condition of limiting equilibrium; with this added empirical information, the problem becomes a determinate one.

3.23.5 The mechanism underlying friction

The origin of frictional forces lies, in the ultimate analysis, in the forces of interaction between those atoms and molecules making up the two bodies in contact that lie in regions close to the area of contact.

It may be mentioned, at the same time, that these electromagnetic forces are principally between the electrons in the atoms or molecules of the two bodies, while the atomic nuclei are also involved in the interaction forces. What is more, the electromagnetic forces between electrons are of a special type since, from the quantum mechanical point of view, electrons belong to the category of fermions (refer to chapter 16 and to section 18.8.9.2 for background).

However, this observation in itself does not explain the characteristic features of frictional forces since the latter depend on a complex interplay of numerous factors that determine their magnitude. In the present introductory exposition, I will touch upon the bare essentials of Bowden and Tabor's theory of dry friction (i.e., friction without lu-

brication) that initiated the modern era of investigations into the mechanism underlying friction and of other significant developments in *tribology*.

The multi-disciplinary science of tribology looks at phenomena involving interacting surfaces in relative motion and is of paramount importance in engineering and technology. It includes in its scope phenomena such as friction, lubrication, abrasion and wear and is of immense economic importance. A renewed interest in hitherto unexplored areas in tribology has erupted in the context of development of devices involving *nanomaterials*.

At the outset, it is to be mentioned that friction is not a material property. It is the property of two surfaces in contact. As in the case of other surface phenomena, the explanation of friction involves complex considerations. And still, a unified and universally accepted theory of friction is yet to emerge, mostly because friction involves a host of distinct phenomena, some of which are of a contingent nature depending on the particular pair of surfaces one is looking at.

A fact of crucial importance to recognize in this context is that even though the surfaces in contact appear to be smooth to the naked eye, there occur, in reality, innumerable microscopic irregularities on these surfaces consisting of small protrusions, termed *asperities*, and valleys or ridges in between these protrusions. As a result of the existence of the asperities, there remain a large number of microscopic *gaps* where the two surfaces are not in actual contact. Fig. 3-69 depicts schematically the nature of actual contact between the surfaces of two bodies, illustrating the distinction between the actual and apparent areas of contact.

Of the atoms and molecules belonging to the two bodies, those that lie close to the actual points of contact, interact with one another by attractive forces when separated by relatively large distances, that give way to strongly repulsive forces at small separations. When a pair of asperities belonging to the two bodies in contact press against each other in virtue of a force normal to the surface of contact, repulsive forces are brought into

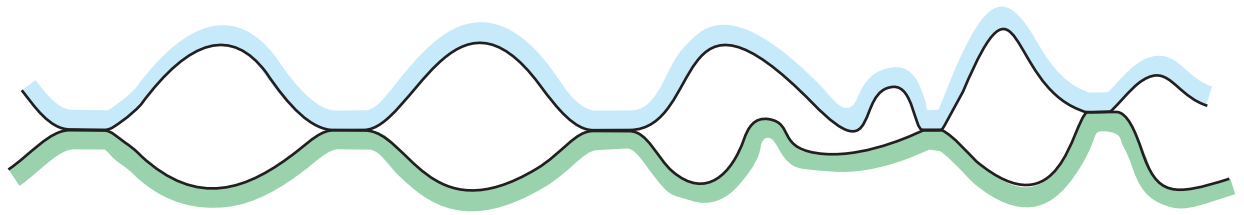


Figure 3-69: Asperities on two surfaces in contact; contact is established between the asperities, due to which the actual area of contact differs from the apparent area, being less than the latter; at the points of contact between the asperities there arise large stresses, associated with elastic and plastic deformations; based on [7], Fig. 5.2.6, p 207.

play, contributing to the normal reaction of one body on the other. If now a force is applied to cause a relative displacement along the surface of separation, local strain fields are produced in the asperities and the shear component of the corresponding stresses result in the generation of the frictional force between the two bodies (see chapter 6 for an introduction to basic concepts relating to *strain* and *stress*). This is illustrated schematically in fig. 3-70(A) and (B) where, in (A), two asperities in actual contact are shown in the absence of an external force or of relative motion, so that there is no strain causing a frictional force (the asperities, however, press against each other producing a compressive stress responsible for the normal reaction). In (B), on the other hand, a shearing strain is produced in each of the asperities as a tendency of relative motion is created (which may or may not result in actual motion), and the associated stress force generated in the asperities is indicated.

The frictional force at such a contact is then the shearing stress times the area of contact over which the stress is developed. Thus, the total frictional force between the two bodies is given by the product of the actual area of contact and the average shearing stress. If the two bodies are not in relative motion, then the contacts between the asperities are not broken and the average shearing stress is less than a certain maximum value, namely, the *yield stress* under shear ($\bar{\tau}$) relevant to the materials of the two bodies (more precisely, the yield stress of the *softer* of the two).

In other words, the force of limiting friction is given by $F = \bar{\tau}A$, where A stands for the *actual* area of contact between the surfaces under consideration.

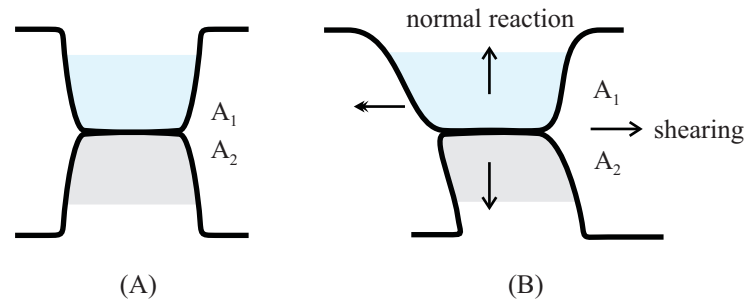


Figure 3-70: Production of shearing stress at the contact between two asperities; (A) asperities A_1 and A_2 are in contact without any tendency of relative motion between the two, as a result of which no shearing stress is developed in either of the two; (B) the double-headed arrow (pointing left) shows direction of applied force or of relative motion; the direction of stress force developed in A_1 is shown (more precisely, imagining a plane in A_1 parallel to the plane of contact, the portion of A_1 lying below the plane exerts an internal force (pointing right) on the portion lying above), the stress force in A_2 being in the opposite direction.

In the case of static friction, as the applied force on any one of the two bodies is increased, causing a tendency of relative motion, the average stress force generated in either body also increases, thereby increasing the frictional force. However, as the average stress force corresponds to the yield stress ($\bar{\tau}$), no further increase of the frictional force is possible. This explains the force of limiting friction which is thus given by the product of the actual area of contact (A) and the yield stress: $F = \bar{\tau} A$.

When there is a relative motion between the bodies, contacts between asperities are continually broken and new contacts established. In this case there is, on the average, a constant shearing stress at each contact close to (but usually somewhat less than) the yield stress. While this average stress during relative motion depends to some extent on the relative velocity one can, in an approximate sense, assume it to be a constant in which case, with other factors (relating to the nature of the two surfaces) remaining unchanged, the frictional force is determined solely by the actual area of contact.

It thus remains to look into the factors that determine the actual area of contact between the asperities. In this, we once again assume that the factors relating to the nature of the two surfaces in contact are given. The extent to which the asperities in the two surfaces come into actual contact then depends on the force of normal action and reaction between the two bodies.

Fig. 3-71(A) shows two asperities before actual contact while 3-71(B) shows the same two asperities after contact. As the two bodies press against each other (due to, say, the weight of one body on the other) local strain fields are once again set up in these, and the associated stress forces are responsible for generating the normal (normal, that is, to the common surface of contact) reaction in the bodies.

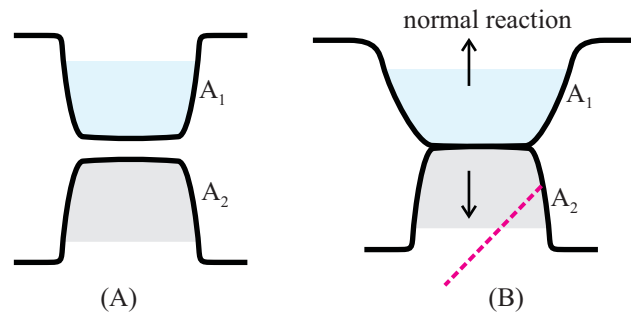


Figure 3-71: Asperities A_1 and A_2 (A) before contact and (B) after contact; stress forces producing normal reaction are shown by arrows; strain fields are developed in the two asperities, including shearing strains in planes such as the one shown by the dotted line in A_2 ; the longitudinal strains are associated with stress forces producing normal reactions, while the shearing strains lead to yielding and plastic deformation.

The deformation in the asperities due to the forces exerted in a direction normal to the common surface of contact may be either *elastic* or *plastic* in nature. The reason why the actual area of contact increases in proportion to the force (N) of normal reaction differs in the two cases.

As the two asperities come into contact and get pressed against each other, they get deformed and the area they present to each other increases. The stress in either of the two materials increases till an *yield stress* is reached.

This development of yield stress involves shearing in the asperities. Thus, there are two distinct causes leading to shearing, one leading to stress forces parallel to the surface of contact and the other producing stress forces normal to the surface. In both these cases, there occurs a slippage between successive parallel planes stacked with atoms in the material in which the shearing takes place (see sec. 6.5.3). Friction can occur between two bodies made of the same material or of different materials. In

the case of bodies made of different materials, the lower of the two yield stresses is of greater relevance in determining the actual area of contact.

While the forces parallel to the surface of contact constitute the friction between the two bodies under consideration, those normal to the surface constitute the *normal reaction* between the bodies.

Since the stress in an asperity at an actual contact cannot exceed the yield stress, and since the total stress force developed in the direction normal to the surface of contact for all the junctions taken together equals the normal reaction (N), the actual area of contact is given, as a first approximation, by the expression $A = \frac{N}{\bar{\sigma}}$, where $\bar{\sigma}$ stands for the yield stress (the stress developed under normal loading for which plastic deformation takes place due to slippage of atomic planes). This explains why the actual area of contact increases in proportion to the force of normal reaction. Taking into account the stress forces parallel and normal to the surface of contact, the force of limiting friction is seen to be $F = \frac{\bar{\tau}}{\bar{\sigma}}N$. Since $\bar{\tau}, \bar{\sigma}$ are characteristics of the two surfaces in contact, one arrives at the Amontons-Coulomb laws of dry friction: the force of limiting friction is independent of the apparent area of contact and is proportional to the normal reaction.

However, friction may not always involve a plastic deformation in the asperities, the latter being the main idea in Bowden and Tabor's theory. In reality, the asperities have a *multi-scale* structure where layers of asperities sit on top of one another, and this is repeated several times over, roughly in what is referred to as a *self-similar* pattern. An increase in the force of normal reaction then involves the creation of *new contacts* rather than an increase of the area at the existing ones in which plastic deformation occurs. In other words, new contacts at a smaller length scale are created where the nature of the deformation at these contacts is elastic rather than plastic.

1. The contact between asperities is intimately related to the statistical features of *surface roughness*. These features include the probability distribution of asperity heights and the correlation between the asperity heights at distinct locations on the surface.

2. The study of stress distributions at elastic contacts between bodies was initiated, in the main, by Hertz, and has since been pursued and developed by several investigators. In particular, investigations into the shearing and tensile stresses developed between rough surfaces in contact constitutes an important area in tribology.

In reality, friction may even involve a combination of both the two types of deformation in the asperities - elastic and plastic.

In summary, frictional forces are produced due to stress forces parallel to the surface of contact as the asperities in actual contact tend to move over or get detached from one another. Assuming a given value of N , the force of normal reaction between the surfaces, the actual area of contact A remains fixed and the total tangential stress force is given by $A\tau$, where τ is the average shearing stress at the asperities. It is this total tangential stress that appears as the frictional force. The maximum value that this force can attain is $A\bar{\tau}$, where $\bar{\tau}$ is the yield stress under shear parallel to the surface. The actual area of contact A increases in proportion to the value of N since it is given by the average number of actual contacts times the average area of each contact. In the case of plastic deformation at the contacts (which occurs as the asperities at a contact are made to press against each other) it is the average area at each contact that increases with N . In the case of elastic deformation, on the other hand, it is the number of contacts that increases. These basic considerations explain a number of observed features relating to friction, including Amontons' law and, in a qualitative manner, Coulomb's law.

While the above outline of the generation of frictional forces relies on the development of elastic strain and stress fields within the asperities, the force of *adhesion* between asperities in contact may also be of direct relevance in determining the frictional force (you will find a brief discussion on this aspect in sec. 3.24.1). The role of the adhesive forces, which may or may not be dominant in determining the force of friction, is qualitatively similar to that of the normal force pressing the two bodies against each other, resulting in similar laws relating to the frictional forces.

Bowden and Tabor's theory is, at times, referred to as the 'adhesion-plasticity' theory of friction, since the forces between the molecules of the two bodies in contact at the tips of the asperities in actual contact result in an adhesive cold-welding of the asperities, and the force of limiting friction depends on the energy required to tear away the adhering tips.

Dynamic (or *kinetic*) friction often involves the mechanism of *stick-and-slip*, especially when the coefficient of dynamic friction is markedly less compared to that of static friction. In moving over one another along the surface of contact, the asperities stick together up to a certain point as in static friction, and then the contacts between the asperities are suddenly broken loose till new contacts are established. In between, the asperities of one of the two bodies slip through the gaps between those of the other body. During this slip phase, elastic vibrations take place in the asperities, similar to those in a spring which is suddenly released after stretching. The energy of these elastic vibrations eventually gets dissipated in the two bodies in contact in the form of heat. In the ultimate analysis, this energy comes from the work done against frictional forces in producing a relative motion between the two bodies in contact. Stick-and-slip is often accompanied by characteristic sounds such as 'chatter' and 'squeak', mostly unpleasant. However, the sound emitted by a bowed string such as in the violin, is also basically due to the stick-slip phenomenon.

Dry kinetic friction often results in the *wear* of material from the two surfaces in contact. As the asperities of the two bodies get deformed on being made to come into contact, layers of material are peeled off from the asperities due to the relative sliding motion of the bodies. Reduction of such wear between moving parts of machines is of great engineering importance, and is effected by means of *lubrication* of the surfaces in contact. As mentioned above, the study of friction, wear, and lubrication, and especially the mechanisms underlying these processes, relates to the subject of tribology (see [7], [8] for an overview of the subject). In recent decades, tribology has progressed in long strides due to the development of a number of techniques allowing the probing of the relevant microscopic processes in great detail.

This brief survey of the mechanisms underlying friction touches upon only a minute fraction of issues relating to dry sliding friction. The mechanism underlying wet friction will be briefly mentioned in sec. 3.23.6. Rolling friction will be taken up, again in bare essentials, in sec. 3.23.7.

Problem 3-56

Problem: Variation of tension along a string under friction.

Consider a weightless string wound on a circular frame as in fig. 3-72(A), where only a part of the winding is shown, covering a length l of the string, the angle subtended by this part at the center of the circular arc being ϕ . If μ be the coefficient of static friction between the string and the frame, and T_1 be the tension at the point A, what is the maximum tension that can be applied at the point B, without the string slipping on the frame?.

Answer to Problem 3-56

Fig. 3-72(B) shows a small part of the string of length δl , with end points P, P', where this element subtends an angle $\delta\theta$ at the center C. Let the tensions at the two ends P and P' be T and $T + \delta T$ respectively. Then, in the limit $\delta\theta \rightarrow 0$, the normal force acting on the element along the normal NC (where N is the mid-point of the arc PP') is $T\delta\theta$ (reason this out) while that along the tangential direction is δT . The maximum possible value of δT in order that the string may not slip on the frame is thus given by the relation $\delta T = \mu T\delta\theta$. This gives the differential equation determining the variation of tension T at any given point (say, P) with the angle θ (see fig. 3-72(A)) as $\frac{dT}{d\theta} = \mu T$. Integrating from $\theta = 0$ (point A, tension T_1) to $\theta = \phi$ (point B, tension T_2), one gets the required value of T_2 as $T_2 = T_1 e^{\mu\phi}$. In other words, the tension increases *exponentially* with the winding angle ϕ (thus, for instance, the difference in the tensions between the ends becomes very large if the string be wound on the frame with multiple turns; the exponential growth in the tension with the angle of winding explains the durability of knots).

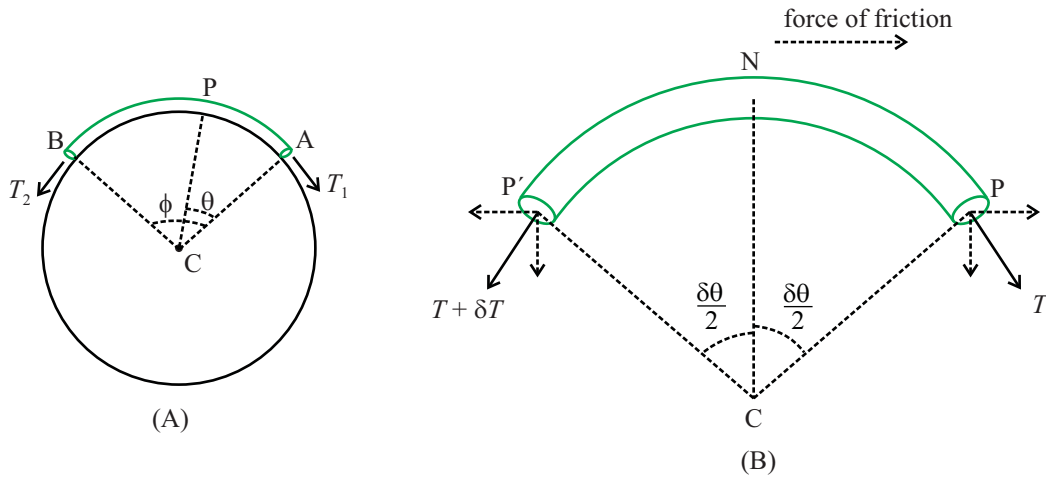


Figure 3-72: A weightless string wound on a rough circular frame; (A) a part of the winding between points A and B is shown, where the circular arc AB subtends an angle ϕ at the center C; (B) a small element P'P of length δl is shown in magnification, which subtends a small angle $\delta\theta$ at the centre C; for a given tension T_1 at the end A of the string in (A), the maximum value of the tension T_2 at the end B can be obtained by referring to the equilibrium of small portions of the string like P'P in (B); for a given value of T_1 , T_2 is seen to increase exponentially with the angle of the winding ϕ ; such exponential dependence on the angle of winding explains the durability of knots.

3.23.6 Wet friction and lubrication

Wet friction is the name given to the frictional resistance to relative motion between two bodies with a fluid film between their surfaces. The presence of the fluid film greatly reduces the frictional force as compared to the force of dry friction. The process of reduction of friction by means of such a fluid film is referred to as *lubrication*.

Broadly speaking, cases of wet friction can be grouped into two classes, namely, *viscous* (or *hydrodynamic*) friction, and *boundary* friction. The distinction between the two is related to the *thickness* of the fluid layer between the two bodies in contact. If the thickness be so large that the asperities of the two surfaces are effectively covered by the fluid layer, i.e., if the asperities do not come into direct contact, then one has a case of viscous friction. In this case, the development of the frictional force is governed by the viscous *drag* (see sec. 7.5.8.3) exerted by the fluid layer on the two bodies. The fluid particles in contact with either of the two surfaces sliding over each other are at rest relative to that surface, and the relative motion between the two surfaces sets up a *velocity gradient* in the fluid layer. The resulting force of viscous drag can be worked

out by making use of the *Navier-Stokes equation* (see sec. 7.5.3) describing the motion of a viscous fluid.

The result of such an exercise shows that the frictional force in viscous friction increases in proportion to the relative velocity between the two bodies under consideration, and is also proportional to the coefficient of viscosity of the fluid. In addition, it also depends on the thickness of the fluid layer, decreasing in magnitude as the thickness increases. One other feature of viscous friction is the *absence of static friction*, which implies that an infinitesimal force applied on either of the two bodies results in a sliding of that body over the other.

If, on the other hand, the thickness of the fluid layer in between the two surfaces in contact is small so that the asperities can come into actual contact, then one has a case of boundary friction. In this case, the characteristic features are similar to those in dry friction, with the major difference that the coefficient of kinetic friction (μ_d) is reduced considerably compared to static friction, and becomes velocity dependent to a greater extent. Commonly, boundary friction is found to arise at low values of the relative velocity when the asperities tend to come into close contact. In practice, one often encounters situations in wet friction where both the viscous and boundary friction mechanisms are involved.

3.23.7 Rolling friction

It is a matter of common observation that the frictional resistance to rolling motion (see sec. 3.20.2) is much less than that in sliding motion. It is this fact that led to the use of wheels and rollers, where it is known that these have been a great ‘driving force’ (in more senses than one!) in history.

Ideally, in the case of pure rolling of one rigid body over another where the instantaneous relative velocity at the point (or line) of contact is zero, there should not arise any resistance at all to motion. However, a small but finite resistance does appear, making a rolling body slow down and causing a conversion of the mechanical energy of motion into thermodynamic internal energy (commonly referred to as ‘heat’) in the two bodies.

In the following we consider the rolling motion of a body B over a flat surface S (referred to as the ‘substrate’) where the materials constituting B and S may possess various different characteristic features relating to their elastic behavior.

If B and S were made of ideally rigid materials and if the contact between the two were at one single point or along one single line, there could be no resistance to rolling motion. In fact, however, the materials get deformed at the contact, generating elastic stress forces over the contact zone. In an ideal situation, the vector sum of these stress forces acting on B just provides the normal reaction (N) necessary to balance its weight, and possesses no component in a direction opposite to that of the motion of B. In practice, however, a small force F_r opposing the motion does appear, and one has to expend energy in the form of work on B so as to make it keep rolling at a constant speed. This energy is eventually dissipated into the materials making up B and the substrate S.

The mechanism underlying the appearance of the force of rolling friction (F_r) and the dissipation of energy can, in principle, be traced to one or more of three causes of distinct nature: (a) interfacial slip between B and S at various points of the deformed contact region, (b) elastic hysteresis, and (c) hysteresis associated with plastic deformation.

A small amount of slipping or sliding between B and S does occur in the contact zone, that can be caused by a difference in the elastic deformations in B and S owing to the difference in their elastic properties, and also by the fact that various different points of contact between B and S distributed over the contact zone lie at varying distances from the instantaneous axis of rotation, as a result of which the relation $v = \omega a$ (see eq. (3-189)) necessary for the absence of sliding is not satisfied at all the points.

However, one finds from experimental observations that the contribution of sliding between the surfaces of B and S to rolling friction is, in most situations of interest, of negligible magnitude. This conclusion is supported by the fact that lubrication does not appreciably affect rolling friction while it reduces sliding friction to a marked extent.

Under certain circumstances, the contribution of *elastic hysteresis* to rolling friction may

be appreciable, where the term ‘elastic hysteresis’ denotes an irreversibility of energy change in a complete cycle of production and release of elastic strain. This can be explained with reference to fig. 3-73, where the deformations produced in S in front of and behind a rolling wheel (B) are shown. As the contact zone in S (which we assume to be made of a softer material compared to B) is deformed, B has to perform work on the region (R_1) of S lying in front of it so as to plow into this region while, at the same time, the strain in the rear region (R_2) is released as the wheel is disengaged from it. This is accompanied by work being done *on* B as the rear region bounces back and pushes on it.

Ideally, the work done *by* B on the front region R_1 should be exactly compensated by the work done *on* it by the rear region R_2 . Expressed in terms of the stress forces on B in these two regions, this means that there should not be any net component of these forces resisting the motion, as mentioned above. In fact, however, the entire amount of the stress energy released in the rear region is not transmitted to the wheel since some of it gets dissipated in the material of the substrate by means of molecular collisions. In other words, there is a net loss of energy in one complete cycle of production and release of elastic strain - a phenomenon referred to as elastic hysteresis which, strictly speaking, is a deviation from ideal elastic behavior of a material.

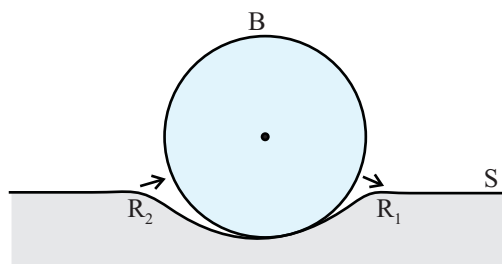


Figure 3-73: Illustrating elastic hysteresis; the deformations in front of and behind the rolling wheel B (regions R_1 and R_2 respectively) are slightly different, where the former corresponds to the wheel pressing on the surface S while the latter to a release of the strain; the work done by the wheel in producing the strain in R_1 is slightly larger than the work done on it as the strain is released in R_2 .

More commonly, however, frictional resistance results from *plastic deformations* pro-

duced in the contact zone, and the hysteresis associated with it. Fig. 3-74 illustrates the phenomenon of plastic hysteresis where the region R_1 in front of the wheel (B) is in the process of being plowed down (where it is eventually to attain the position shown by the dotted line) while the rear region R_2 is in the process of bouncing back, but only to a partial extent. The line L_1 denotes the original position of this region before it was plowed down by B to the position L_2 , from where it bounces back to the position L_3 . Evidently, the energy expended on R_1 as it is deformed by B, is not the same as the energy released by R_2 as it undergoes the process of incomplete bounce-back, and a net energy, equal to the difference between the two, gets dissipated in S (in practice, there occurs deformation and dissipation in B as well). In terms of stress forces developed over the contact zone, there results a net force F_r opposing the motion.

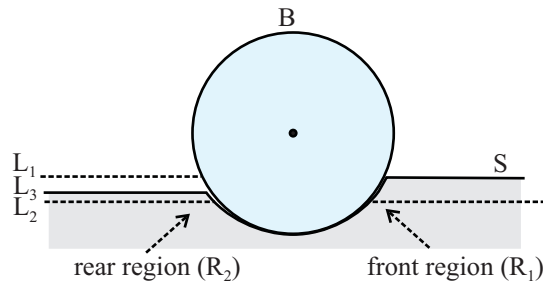


Figure 3-74: Illustrating plastic hysteresis in rolling motion of the wheel B on the surface S; the region R_1 in front of the wheel is in the process of being plowed down to the level shown by the dotted line, while the strain in the rear region R_2 is being released from the level L_2 to L_3 , which differs from the original level L_1 (schematic).

From the point of view of the production of a resistance to rolling motion, the effects of elastic and plastic hysteresis are essentially similar, the only difference being in the magnitudes of the resisting forces which relates, in the ultimate analysis, to the internal mechanisms responsible for the development and release of stresses in the materials concerned (see sec. 6.5.3).

Fig. 3-75 depicts schematically the resultant stress force F operating on the rolling body B (say, a circular wheel) due to the stress forces distributed over the contact zone, where we assume that the only other external force acting on B is its weight W . The horizontal

component F_r of the resultant stress force is the force of rolling friction slowing down the rolling motion, while the vertical component N provides the normal reaction balancing the weight of B. In practice, the angle θ between the line of action of F and the vertical direction is so small that one has $F \approx N = W$, and

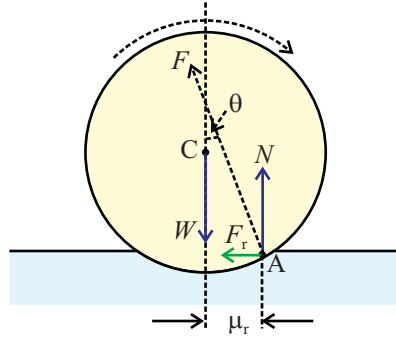


Figure 3-75: Depicting the direction of the reaction force F in rolling friction, the horizontal component of which gives the frictional force F_r ; the vertical component N gives the normal reaction, balancing the weight W ; θ denotes the inclination of the reaction force F to the vertical; F is the resultant of contact forces and acts effectively through the point A in the contact zone; there is a slight asymmetry (not shown in the figure) between the deformations in the front and rear zones due to elastic or plastic hysteresis, causing the reaction force as shown.

$$F_r = \mu_r \frac{N}{R}, \quad (3-211)$$

where R stands for the radius of the wheel and μ_r is distance shown in the figure, referred to as the coefficient of rolling friction which, in contrast to the coefficient of kinetic friction, has the dimension of length. It can be defined as the distance of the point A on the deformed surface of contact (through which the stress force acts) from the vertical line passing through the center of the wheel. The reaction force $F(\approx N)$ produces a torque about the center opposing the rotational motion, and is referred to as the frictional torque, which is given by

$$\tau = \mu_r N. \quad (3-212)$$

The relations (3-211) and (3-212) express the phenomenological law of rolling friction,

referred to as *Coulomb's law*. Strictly speaking, Coulomb's law of rolling friction states that the coefficient μ_r in the above equations is independent of the velocity of rolling. In reality, however, μ_r may depend to some extent on the velocity.

3.24 Motion of a wheel under driving

The concept of pure rolling (or, in brief, simply *rolling*), as distinct from rolling with sliding, was introduced in sec 3.20.2, while rolling friction was considered in sec. 3.23.7, where we saw that rolling friction, while much smaller in magnitude compared to sliding friction, is involved in the slowing down of a freely rolling wheel.

In the present section we consider the motion of a wheel under the action of an externally applied force or an external torque, i.e., under an external driving. Such external driving is employed in wheeled vehicles such as an automobile. For the sake of simplicity, we shall assume that the force of rolling friction is negligibly small. The force of static sliding friction, however, plays an essential role in the motion of the wheel.

While an automobile is a *self-propelled* vehicle, its engine acts as an external driving agent with respect to its wheels - commonly, the rear wheels, on which a torque is applied by means of mechanical coupling with the engine. The rear wheels, in turn, apply a force on the front wheels, imparting a rolling motion. Thus, the driving of the rear wheels takes place by means of a torque while that of the front wheels is by means of a force applied to their center.

3.24.1 Driving by means of a couple

Fig. 3-76 depicts a wheel of radius R where we assume for the sake of simplicity that the plane of the wheel is vertical and where a driving couple M is assumed to act on the wheel in its plane in the sense shown. The wheel rolls on a horizontal plane with an angular velocity ω , the velocity of the center of the wheel being v .

Since the couple tends to rotate the wheel in the sense shown by the bent arrow, a frictional force F_f acts on it in the direction shown, and it is this force, exerted by the

ground, that is responsible for the forward motion (from the left to the right in the figure) of the wheel. Since we assume that the motion of the wheel is one of pure rolling, the frictional force is due to static friction (or, in part, due to *adhesion*) between the ground and the wheel.

In general, the wheel may possess an acceleration, say, a (along a direction from the left to the right in the figure), and an angular acceleration, say α (in the clockwise sense) where

$$F_f = ma, \quad (3-213a)$$

$$M - F_f R = I\alpha, \quad (3-213b)$$

m and I being respectively the mass of the wheel and its moment of inertia about an axis perpendicular to its plane and passing through its center (assumed to be its center of mass).

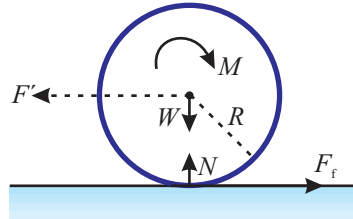


Figure 3-76: Driven motion of a wheel of radius R under a couple M ; the frictional force F_f acts on the wheel in the forward direction; the wheel accelerates in the forward direction; for the wheel to eventually attain a uniform velocity, a force F' is to act on it in the opposite direction; in practice this may arise from a velocity-dependent drag due to air resistance; W and N are the weight of the wheel and the force of normal reaction on it.

Eq. (3-213a) describes the center of mass motion of the wheel, while (3-213b) describes its angular motion about the center of mass. The acceleration a and the angular acceleration α appearing in these equations are the rates of change of v and ω respectively,

where the condition of pure rolling is (refer to eq. (3-189))

$$v = \omega R. \quad (3-214)$$

Thus, assuming that the wheel keeps on rolling without sliding on the ground (the condition for this will be derived below), one gets

$$F_f = \frac{mMR}{I + mR^2}. \quad (3-215)$$

Evidently, this force has to be less than or equal to the maximum force of static friction that the ground can apply on the wheel, which is given by μN , where μ is the relevant coefficient of friction and N is the force of normal reaction which, in the present instance, equals the weight of the wheel (in the case of the rear wheel of an automobile, N includes a part of the total weight of the body of the car).

Thus, in other words, the wheel moves with a uniform acceleration a (and a corresponding angular acceleration $\alpha = \frac{a}{R}$) if

$$\frac{MR}{R^2 + k^2} < \mu N, \quad (3-216)$$

during which a frictional force given by eq. (3-215) (i.e., the left hand side of (3-216)) acts on it, which is obtained by writing $I = mk^2$, k being the radius of gyration corresponding to the moment of inertia I . With this value of the frictional force, the acceleration a of the wheel is given by the eq. (3-213a)

In some situations, F_f may, strictly speaking, be determined by the *adhesive* force between the ground and the wheel. While the mechanism underlying the generation of the frictional force between two bodies relates to the elastic properties of the materials of the bodies, the adhesive force depends on the forces of interaction between the molecules of the two bodies lying close to the region of contact. Though we have termed F_f the force of static friction, it may arise mainly due to the adhesive force between the ground and the wheel. While this adhesive force is of negligible importance in

numerous situations of interest, it may be the dominant one in some others. The adhesive force, like the force of static friction, can have a maximum value μN , where μ is a constant characterizing the adhesive interaction. In considering friction, one need not, in general, distinguish between the adhesive force and the frictional force determined by the elastic properties of the materials in contact. In some situations, the adhesive force may be relevant in determining the force of rolling friction.

In practice, the wheel cannot accelerate to an infinitely large velocity. In the case of an automobile, for instance, there arises a velocity-dependent force on the body of the automobile due to the viscous drag exerted by the surrounding air, which operates in a direction opposite to the motion, and this force is transmitted to the wheels. As the velocity increases, the drag force also increases till at a certain velocity it balances the force F_f on the wheel.

Thus, assuming that a horizontal velocity-dependent force $F'(v)$ acts on the wheel in the negative direction, eq. (3-213a) gets modified to

$$F_f - F'(v) = Ma. \quad (3-217)$$

The terminal velocity v_0 of the wheel is given by the condition

$$F'(v_0) = F_f = \frac{M}{R}. \quad (3-218)$$

In other words, the wheel will perform a motion of pure rolling and will accelerate up to a velocity v_0 given by $F'(v_0) = \frac{M}{R}$ (recall that M is the driving couple applied on the wheel), when the frictional force F_f on it is exactly balanced by the resistive force F' . The condition that such a motion will actually occur is given by

$$\frac{M}{R} < \mu N. \quad (3-219)$$

If the wheel starts from rest, the resistive force F' is zero, and the condition of pure rolling is given by the inequality (3-216). As the wheel attains its terminal velocity,

however, the stronger condition (3-219) becomes necessary. In other words, even if sliding is prevented at the beginning, the motion may cease to be one of pure rolling as the wheel approaches its terminal velocity.

I repeat that, in these considerations, we have neglected the force of rolling friction on the wheel.

3.24.2 Driving by means of a force

Fig. 3-77 depicts a wheel as in fig. 3-76, but now under the driving by a force F applied horizontally through its center instead of by a couple. As indicated in sec 3.24.1, a resistive force $F'(v)$ may, in general act on the wheel, where $F'(v)$ is an increasing function of v . The equations governing the motion of the wheel are now

$$F - F_f - F' = ma, \quad (3-220a)$$

$$F_f R = I\alpha, \quad (3-220b)$$

where, in the case of pure rolling, a and α are once again related by $a = \alpha R$. Comparing fig. 3-77 with 3-76, one observes that, in the case of driving by means of a force F applied to the center of the wheel, the frictional force F_f is in the *negative* direction since the applied force tends to push the wheel over the ground in the forward (positive) direction. At the initial stage, if the wheel starts from rest ($F' = 0$), the frictional force necessary for pure rolling works out to

$$F_f = \frac{F}{1 + \frac{R^2}{k^2}}, \quad (3-221)$$

where k is once again the radius of gyration corresponding to I . Accordingly, the motion will actually be one of pure rolling if the applied force satisfies

$$\frac{F}{1 + \frac{R^2}{k^2}} < \mu N. \quad (3-222)$$

The wheel accelerates till it attains a limiting velocity v_0 when $\alpha = 0$, i.e., the frictional force F_f gets reduced to zero. In this case, v_0 is determined by

$$F'(v_0) = F. \quad (3-223)$$

Since the frictional force reduces to zero as the terminal velocity is achieved, the condition (3-222) necessary for the motion to be one of pure rolling at the beginning of the motion suffices for the motion to be a pure rolling till the end when the terminal velocity is reached, in contrast to the case of driving by means of a couple.

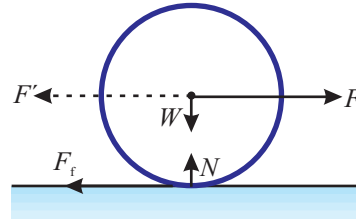


Figure 3-77: Depicting the rolling motion of a wheel under a driving force F applied horizontally to the center; the frictional force F_f acts in the backward direction; for rolling with uniform velocity an oppositely directed force F' is necessary; W and N denote the weight of the wheel and the force of normal reaction on it.

In order to slow down an automobile, the driver applies the brakes to the wheels, whereby pad-like objects (the *brake-shoes*) are made to press against the body of the wheels. The resulting forces of sliding friction cause a couple M_b to act on the wheels in a direction opposite to the driving. Thus, in eq. (3-218), the driving moment M is to be replaced with $M - M_b$, thereby implying that v_0 is reduced.

Problem 3-57

Consider a wheel in the form of a homogeneous disk of mass $m = 0.5$ kg and radius $R = 0.5$ m, undergoing pure rolling on horizontal ground, with the plane of the disc remaining vertical, where the wheel is driven by a force 1 N applied horizontally through its center, and a couple of moment 1 N·m acting in the plane of the disc (the direction of the force and the sense of the couple being

as in figures 3-77 and 3-76 respectively). Find the frictional force at the point of contact and the acceleration of the wheel. What is the minimum value of the coefficient of friction between the wheel and the ground for sliding to be prevented ($g = 9.8 \text{ m}\cdot\text{s}^{-2}$)?

Answer to Problem 3-57

Here the moment ($M = 1$) (all SI units implied) of the applied couple and the applied force ($F = 1$) are related as $M > \frac{I}{mR} F$ ($I = \frac{1}{2}mR^2$), and hence the force of friction (F_f) on the wheel at the point of contact acts in the forward direction (reason this out; however, see below).

One has, $M - F_f R = I\alpha$, and $F + F_f = ma$, where $I = \frac{1}{2}mR^2$ ($m = 0.5$), and a and α are the linear and angular accelerations, related by $a = \alpha R$ in the case of pure rolling. This gives $a = 4 \text{ m}\cdot\text{s}^{-2}$. The frictional force on the wheel works out to $F_f = 1 \text{ N}$. Since this has to be less than $\mu W = \mu mg = 4.9\mu \text{ N}$, the coefficient of friction must satisfy $\mu > \frac{1}{4.9}$.

The fact that F_f is positive is seen to result from our analysis, and is not really needed as a pre-requisite for the same. Had we obtained a negative value of F_f , then that would have meant that the frictional force is in the negative direction for the given values of the parameters. For the special case $M = \frac{I}{mR} F$, one would have obtained $F_f = 0$.

Problem 3-58

A homogeneous circular disk of radius r and mass m rolls without slipping on a rough horizontal plane with a velocity v , with its plane perpendicular to the boundary line where the horizontal plane changes to an inclined plane whose angle of dip is α (fig. 3-78). Determine the maximum value of v for which the disc rolls on to the inclined plane without being separated from the supporting surface.

Answer to Problem 3-58

HINT: As the disk rolls on to the incline without getting detached from the edge of the plane, its center C moves in a circular arc (bent arrow in fig. 3-78) about the point P. If, at any instant of time, the angle of rotation of the radial line PC from the vertical is θ , then the vertical height descended by the center (the center of mass of the disk in this case) is $r(1 - \cos\theta)$. The energy balance equation is then $\frac{3}{4}mv^2 = \frac{3}{4}mv'^2 - mgr(1 - \cos\theta)$, where v' stands for the velocity of the

center of mass in the position shown in the figure (the factor of $\frac{3}{4}$ appears, instead of $\frac{1}{2}$, because of the fact that the instantaneous motion of the disk is one of rotation about an axis perpendicular to its plane and passing through a point on its periphery). The centripetal acceleration of the center of the disk is accounted for by the resolved component of the weight mg along CP, i.e., $mg \cos \theta = \frac{mv'^2}{r}$, assuming the reaction force on the disk to be zero.

These two relations give $\cos \theta = \frac{3v^2}{7gr} + \frac{4}{7}$. The radial line CP is to rotate by $\theta = \alpha$ if the disk is to roll on to the incline. For a velocity v larger than a limiting value v_0 , the resolved component of the weight will not be able to account for the centripetal acceleration as θ increases to α while, for a lower velocity, there remains a reaction force on the disk at the point P along PC that enters into the force equation; for $v = v_0$ the reaction force drops to zero when $\theta = \alpha$. One therefore obtains the limiting velocity

$$v_0 = \sqrt{gr\left(\frac{7}{3} \cos \alpha - \frac{4}{3}\right)}. \quad (3-224)$$

NOTE: This means that, for $\alpha \geq \arccos \frac{4}{7}$, the disk will get detached *regardless* of its initial velocity v .

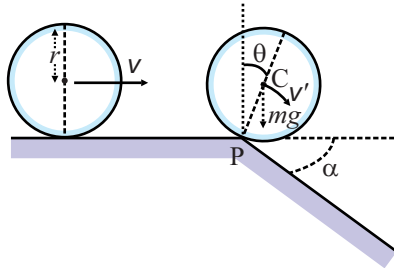


Figure 3-78: A homogeneous circular disk of radius r and mass m rolls without slipping on a rough horizontal plane with a velocity v , with its plane vertical and perpendicular to the boundary line where the horizontal plane changes to an inclined plane whose angle of dip is α ; the maximum value of v for which the disc rolls on to the inclined plane without being separated from the supporting surface is given by formula (3-224).

Chapter 4

Simple Harmonic Motion

The principles underlying the dynamics of a particle introduced in chapter 3, will be made use of in the present chapter to describe and explain a type of motion of great importance in physics, referred to as *simple harmonic motion*.

4.1 Oscillatory motion

Imagine a light and rigid rod to be fixed at its upper end (S), with a small heavy mass (the 'bob') attached to its lower end, as in fig. 4-1(A). If the bob is now pulled through a small distance on one side of the equilibrium position O, say, to A, and is released from rest, then it will be found that the rod along with the bob keeps on swinging on two sides of the mean position SO, and the bob moves to and from A to A' and back. This is an instance of a *periodic* motion, where the state of motion of the system is repeated at fixed intervals of time. The motion of the system is referred to as an *oscillation*. The motion of the bob is sometimes also referred to as a *vibration*. The system made up of the rod and the bob is called a *simple pendulum*.

The terms 'oscillation' and 'vibration' usually carry different connotations, but are sometimes used interchangeably in order to refer to periodic motions. Periodic motions other than oscillatory ones are also possible like, for instance, *rotation*.

4.1.1 Simple harmonic motion

4.1.1.1 The equation of motion

In the above example, if the displacement of the bob on either side of the equilibrium position (i.e., the distance from O to A or A') be small and if the bob be so small in size that it can be considered to be a point mass then in an approximate sense the motion of that point mass about its mean position O can be described as an oscillatory motion along a straight line. The force on the bob at any position during its motion is found to be (see below) directed toward the mean position O, its magnitude being proportional to the instantaneous displacement from O. Such a force is said to be a *linear restoring* one, and the periodic oscillation of a particle under such a force is termed a *simple harmonic motion* (SHM in brief).

One thus has the following definition: the motion of a point mass along a straight line under a force whose magnitude is proportional to the instantaneous displacement of the particle from a fixed point (say, O) on the line and which is always directed toward that fixed point, is termed a simple harmonic motion (or, in brief, SHM).

In fig. 4-1(B), P denotes the instantaneous position of the particle along the straight line X'OX, which we take to be the x-axis with origin at O, the mean position of the particle. The displacement measured from the origin, velocity, acceleration, and force are all taken to be positive when directed towards the positive direction of the x-axis, i.e., from left to right while their signs are taken to be negative when these are directed in the opposite direction. Then, according to the above definition, the instantaneous restoring force on the particle (sometimes referred to as the 'force of restitution') for a displacement x is given by the expression $-kx$, where k , the force per unit displacement, is a constant referred to as the force constant for the motion.

The negative sign in the expression of the force is significant. If the displacement x is positive, i.e., the position of the particle is to the right of O, then the above expression tells us that the force is negative, i.e., directed towards the left or, in other words,

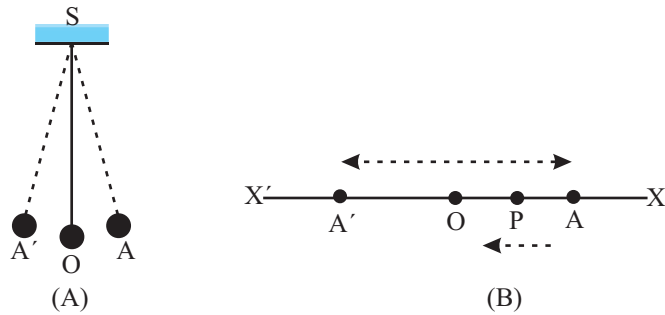


Figure 4-1: (A) A simple pendulum with a light rigid rod fixed at the upper end (S) and a mass attached to the lower end; O is the equilibrium position of the mass, while A and A' are the two extreme positions; the bob oscillates between A and A'; (B) simple harmonic motion executed by a particle along X'OX; O is the mean position while P is the instantaneous position at any given point of time; A and A' are the two extreme positions; the to-and-fro motion is indicated with the two-way arrow; the direction of force in the position P is indicated with a dashed arrow.

toward the mean position O. If, on the other hand, the displacement is toward the left of O, then the force is positive, i.e., directed towards the right, once again toward O. It is this inverse relation between the directions of displacement and force that is expressed by the negative sign in the above expression of the force.

Using the equation of motion for a particle in one dimension (eq. (3-29b)), and the above form of the force on the particle, one gets the following *equation of motion* for the particle in SHM:

$$m \frac{d^2x}{dt^2} = -kx. \quad (4-1)$$

The value of k for the case of the simple pendulum described above can be worked out by considering the forces acting on the bob at any point of time in its oscillatory motion. For sufficiently small amplitude (i.e., maximum displacement from the mean position) of oscillation, the net horizontal component of the force works out to $-m \frac{g}{l} x$, where g stands for the acceleration due to gravity and l for the length of the rod. This gives $k = \frac{mg}{l}$.

4.1.1.2 Solving the equation of motion

This is a second order differential equation. Solving this equation, one gets to know how the position co-ordinate (x) and the velocity ($v = \dot{x}$, where the dot denotes differentiation with respect to time) of the particle varies with time (t). For this, let us set

$$\sqrt{\frac{k}{m}} = \omega, \quad (4-2)$$

and multiply both sides of eq. (4-2) with $2\dot{x}$. This gives

$$\frac{d}{dt}(\dot{x}^2) = -\omega^2 \frac{d}{dt}(x^2), \quad (4-3)$$

which gives, on integration,

$$\dot{x}^2 + \omega^2 x^2 = C_1, \quad (4-4)$$

where C_1 is a constant of integration. As we will see (refer to eq. (4-25) below), this first integral of the equation of motion, i.e., the result of performing an integration on it, is nothing but the equation expressing the conservation of energy for the system under consideration.

Noting that the constant of integration C_1 in eq. (4-4) is necessarily positive and satisfies $C_1 \geq \omega^2 x^2$, one can perform a second integration now (recall that the equation of motion is a second order differential equation) to obtain

$$\int \frac{dx}{\sqrt{(C_1 - \omega^2 x^2)}} = t + C_2, \quad (4-5)$$

where C_2 is a second constant of integration. Working out the indefinite integral on the left hand side, one obtains

$$\frac{\omega x}{\sqrt{C_1}} = \cos(\omega(t + C_2)), \quad (4-6)$$

which can be written as

$$x = a \cos(\omega t + \delta), \quad (4-7)$$

where a and δ are two new constants related to C_1 and C_2 .

Note, incidentally, that the ambiguity of sign in taking the square root in eq. (4-5) does not affect the final result (4-7) one arrives at.

4.1.1.3 General and particular solutions

The significance of the constants a and δ in respect of the motion of the particle under consideration will be indicated below. Eq. (4-7), containing the two undetermined constants a and δ (or, equivalently, C_1 and C_2), constitutes the *general solution* of the equation of motion, eq. (4-1), of the particle. Considering any particular instance of motion, with given initial conditions, one can express the constants in terms of the initial position and velocity of the particle, as we will see below.

One can express the solution in another equivalent form as

$$x(t) = A \cos \omega t + B \sin \omega t. \quad (4-8)$$

where A and B are two new constants, independent of time, related to the pair a, δ as

$$A = a \cos \delta, \quad B = -a \sin \delta, \quad (4-9a)$$

or, equivalently, as

$$a = \sqrt{A^2 + B^2}, \quad \delta = \arctan\left(-\frac{B}{A}\right). \quad (4-9b)$$

The parameter ω characterizing the motion is referred to as its *angular frequency*, and determines the *time period* of the motion (see below).

Assuming that the values of the constants A, B (or, equivalently, of a, δ) are known, one

gets a *particular solution* of the equation of motion. Else, if these are treated as undetermined constants then, as mentioned above, (4-8) constitutes the general solution of the equation. Along with the position co-ordinate of the particle at time t , the solution (4-8) also gives the *velocity* at time t :

$$v = \frac{dx}{dt} = -\omega A \sin \omega t + \omega B \cos \omega t. \quad (4-10)$$

4.1.1.4 Relating the solution to initial conditions

If one knows the position and velocity of the particle at any particular instant of time, say, at $t = 0$, then one can work out the values of the constants A and B in (4-8) with the help of these data, thereby arriving at a particular solution of the equation of motion. In the context of the particular solution, the data consisting of the position and velocity at time $t = 0$ (or at any other given time) are referred to as the *initial conditions*.

As an example, let the position and velocity at time $t = 0$ be x_0 and v_0 respectively. Then, making use of (4-8), (4-10) one gets

$$A = x_0, \quad B = \frac{v_0}{\omega}, \quad (4-11)$$

and the particular solution for these initial conditions is seen to be

$$x = x_0 \cos \omega t + \frac{v_0}{\omega} \sin \omega t, \quad v = -\omega x_0 \sin \omega t + v_0 \cos \omega t. \quad (4-12)$$

The term 'initial' is used above to designate any particular chosen instant of time.

If this instant were chosen to be, say, t_0 rather than 0, then the particular solution obtained from initial conditions $x = x_0, v = v_0$, at $t = t_0$ would be

$$x = x_0 \cos(\omega(t - t_0)) + \frac{v_0}{\omega} \sin(\omega(t - t_0)), \quad v = -\omega x_0 \sin(\omega(t - t_0)) + v_0 \cos(\omega(t - t_0)). \quad (4-13)$$

(check this out).

Using the form (4-7) of the solution, one obtains the following expression for the velocity of the particle at any instant t

$$v = -a\omega \sin(\omega t + \delta), \quad (4-14)$$

which, along with (4-7) implies that the velocity at position x can be obtained from the expression

$$v^2 = \omega^2(a^2 - x^2). \quad (4-15)$$

(check this out).

For any given position x (note that x varies periodically in the range $-a \leq x \leq +a$) there are two solutions to the velocity v in eq. (4-15), the two velocities being equal in magnitude and opposite in direction. This corresponds to the fact that the particle passes through any given point twice in each period (i.e., say, in the interval $0 \leq t \leq T$, see sec. 4.1.1.5) in opposite directions, i.e., referring to fig. 4-1(B), once from left to right and then again from right to left.

Notice that, while determining the value of a in terms of A and B , the square root of $(A^2 + B^2)$ can be taken with either a positive or a negative sign. One commonly chooses the positive sign for the sake of definiteness. However, even after the choice of a definite sign for a , there remains a non-uniqueness in the value of δ since adding 2π to any possible value gives another acceptable value of δ . In order to eliminate this non-uniqueness, one usually chooses δ in the range $0 \leq \delta < 2\pi$, or $-\pi \leq \delta < \pi$.

4.1.1.5 Periodicity of motion

Equations (4-7), (4-14) tell us that a simple harmonic motion is indeed a periodic motion since the trigonometric functions \sin and \cos (these are also referred to as *circular functions*) are periodic in their arguments. For instance, one has

$$\sin(\theta + 2n\pi) = \sin \theta, \quad \cos(\theta + 2n\pi) = \cos \theta, \quad (n = \pm 1, \pm 2, \dots). \quad (4-16)$$

This means that substituting $t + \frac{2\pi}{\omega}$ in place of t in (4-7), (4-14) (or in any equivalent form of the solution to the equation of motion), the argument of the sin and cos functions gets changed from ωt to $(\omega t + 2\pi)$, giving back the values of displacement and velocity one started with at time t :

$$x(t + \frac{2\pi}{\omega}) = x(t), \quad v(t + \frac{2\pi}{\omega}) = v(t). \quad (4-17)$$

In other words, writing

$$T = \frac{2\pi}{\omega}, \quad (4-18)$$

one concludes that the state of motion (i.e., the displacement *along with* the velocity) is repeated at intervals of temporal extension T , referred to as the *time period* of oscillation. One notes that the state of motion at any time t is repeated at times $t + T, t + 2T, \dots$, but not at any time from, say, t to $t + T$. This means that the time period is the *minimum* interval of time after which the state of motion is repeated.

Corresponding to the time period T , one can define the *frequency* ν as the reciprocal of the time period, which stands for the number of times the state of motion gets repeated in a unit time interval:

$$\nu = \frac{1}{T} = \frac{\omega}{2\pi}. \quad (4-19)$$

4.1.1.6 The phase

The time-dependence of the state of motion in equations (4-7), (4-14) is seen to occur through the expression $\omega t + \delta$. Denoting this by $\Phi(t)$, one can write, in brief,

$$x(t) = a \cos \Phi(t), \quad v(t) = -a\omega \sin \Phi(t), \quad (4-20)$$

where $\Phi(t)$ is referred to as the *phase angle* (or, simply, the *phase*) of the motion.

The reason why $\Phi(t)$ is termed the phase *angle* is that it occurs as the argument of the trigonometric functions sin and cos in (4-20), and hence the values of $x(t)$, $v(t)$ remain

unchanged if any positive or negative integral multiple of 2π is added to $\Phi(t)$. In other words, only the value of Φ *modulo* 2π is relevant in determining the state of motion at any given instant of time. In course of time, as t keeps on increasing, $\Phi(t)$ also increases monotonically, but the value of $\Phi(t)$ modulo 2π gets repeated periodically.

The phrase ' $\Phi(t)$ modulo 2π ' means an angle (let us call it $\phi(t)$, the *reduced phase*) lying in the range $0 \leq \phi(t) < 2\pi$, obtained by adding an appropriate (positive or negative) multiple of 2π to $\Phi(t)$. Alternatively, one can choose the reduced phase to lie in the range $-\pi \leq \phi(t) < \pi$. It is usually not difficult to pick up from the context whether it is $\Phi(t)$ or the reduced phase $\phi(t)$ that is being referred to. The term 'phase' is commonly used to denote either $\Phi(t)$ and $\phi(t)$ as the context demands.

According to the definition of the phase, one finds that δ is nothing but the value of the phase at time $t = 0$. It is therefore referred to as the *initial phase* of the motion.

4.1.1.7 The amplitude

Looking now at the constant a in eq. (4-7), one observes that $x(t)$ oscillates in course of time between the values $-a$ and a , i.e., its maximum displacement on either side of the mean position $x = 0$ is a . This is referred to as the *amplitude* of the simple harmonic motion. In a similar manner, the velocity oscillates between $-a\omega$ and $a\omega$, where a negative sign of the velocity means that the instantaneous direction of motion is towards the negative side of the x-axis. Note, moreover, that the velocity is zero as the particle reaches either of its two extreme positions, i.e., for $x = -a$ and $x = a$.

In order to see how the periodic changes in the displacement and velocity are related to each other, it is useful to rewrite the expression of velocity in terms of the \cos function, i.e., by writing it is $v(t) = a\omega \cos(\Phi(t) + \frac{\pi}{2})$. Comparing this with the expression for $x(t)$ in eq. (4-20), one notes that, while the \cos function is used in both expressions, the *arguments* differ in the two. If $\Phi(t)$ is referred to as the phase angle of the displacement, then the phase angle of the velocity has to be identified as $\Phi(t) + \frac{\pi}{2}$. One expresses this by saying that the phase angle of velocity *leads* that of the displacement by $\frac{\pi}{2}$.

Alternatively, one says that the phase of the displacement *lags* that of the velocity by $\frac{\pi}{2}$.

Problem 4-1

A particle of mass 0.01 kg executes a simple harmonic motion where the restoring force per unit displacement is $0.1 \text{ N}\cdot\text{m}^{-1}$. What is the time period of oscillation? If the initial displacement of the particle from the mean position is -0.4 m and the initial velocity is $1.0 \text{ m}\cdot\text{s}^{-1}$, find the amplitude of oscillation and the displacement and velocity at $t = 30.0 \text{ s}$.

Answer to Problem 4-1

HINT: $T = \frac{2\pi}{\sqrt{\frac{k}{m}}}$, where (all SI units implied) $m = 0.01$, $k = 0.1$ (thus, $T = 1.99 \text{ s}$ (approx)). Choosing the expression for displacement in the form $x = a \cos(\omega t + \delta)$, one has $-0.4 = a \cos \delta$, $1.0 = -a\omega \sin \delta$, where $\omega = \sqrt{\frac{0.1}{0.01}}$. This gives $a^2 = (0.4)^2 + \frac{(1.0)^2}{10} = 0.26$, i.e., $a = 0.501 \text{ m}$, and $\delta = \tan^{-1} \frac{1.0}{0.4 \times \omega}$, where δ lies in the range $\pi \leq \delta \leq \frac{3\pi}{2}$, i.e., $\delta = 3.809$. These values can now be used to determine $x = a \cos(\omega t + \delta)$, and $v = -a\omega \sin(\omega t + \delta)$, with $t = 30.0$. One obtains $x = -0.14 \text{ m}$, $v = 1.55 \text{ m}\cdot\text{s}^{-1}$.

Problem 4-2

A particle of mass 0.2 kg oscillates along a straight line, bouncing successively between two fixed walls perpendicular to the line of motion (see fig. 4-2), the collisions with the walls being elastic. In between the walls, the particle is acted upon by a restoring force directed towards the mid-point between the walls, where the restoring force per unit displacement is $1.0 \text{ N}\cdot\text{m}^{-1}$. If the distance between the walls is $2.0 \text{ m}\cdot\text{s}^{-1}$, and the total energy of the particle is 4.5 J , find the time period of motion. Is the motion sinusoidal (i.e., one where the displacement can be expressed in the form of a sine or cosine function)?

Answer to Problem 4-2

HINT: If the walls were not there, the particle would execute a simple harmonic motion with an amplitude a where $E = \frac{1}{2}m\omega^2 a^2 = \frac{1}{2}ka^2$, where (all SI units implied) $E = 4.5$, $k = 1.0$, i.e., $a = 3.0$. Since $2a > d (= 2.0)$, the distance between the walls, the particle bounces off the walls without completing its full swing under the restoring force, and the motion is non-sinusoidal, though periodic. The velocity v_0 at the instant of a collision is obtained from the equation $\frac{1}{2}mv_0^2 + \frac{1}{2}kx^2 =$

4.5, where $m = 0.2$, $k = 1.0$, $x = \frac{2.0}{2}$ (thus $v_0 = 6.325$). The velocity gets reversed in an elastic collision. Starting from the instant, say, $t = 0$ when $x = a \cos \delta = -1.0$, and $v = -a\omega \sin \delta = 6.325$ (immediately after a collision with the wall on the left), δ is obtained by putting $a = 3.0$, $\omega = \sqrt{\frac{1.0}{0.2}}$ (this gives $\delta = \pi + 1.230$). The time τ taken to reach the opposite wall is then given by $1.0 = 3.0 \cos(\omega\tau + \delta)$, from which one can determine the least value of τ , which is seen to be $\tau = 0.304$. The time period of oscillation is then $2\tau = 0.608$ s.

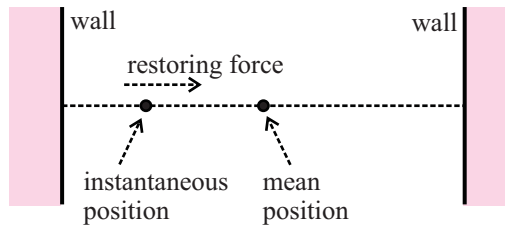


Figure 4-2: A particle moving between two fixed walls, bouncing elastically from each of the two and experiencing a restoring force in between, the latter being directed towards the mid-point between the walls; if the amplitude of the simple harmonic motion caused by the restoring force is less than half the distance between the walls, the particle does not suffer collisions with the walls; for a larger value of the amplitude, the simple harmonic motion is punctuated with successive collisions.

4.1.1.8 Graphical representation of the motion

The way the displacement $x(t)$ and the velocity $v(t)$ change periodically with time can be depicted graphically by plotting these two against time t . This is shown in fig. 4-3, where I have, for the sake of definiteness, taken the initial conditions as $x(t = 0) = x_0 (> 0)$, and $v(t = 0) = 0$, which means that $x(t)$ and $v(t)$ are given by

$$x(t) = x_0 \cos \omega t, \quad v(t) = -x_0 \omega \sin \omega t, \quad (4-21)$$

and that the amplitude of oscillations in $x(t)$ is $a = x_0$.

As seen from the graphs, the state of motion, i.e., the displacement along with the velocity, is repeated at intervals of time T , where, during each such interval, the phase increases by 2π . But as mentioned above, a change in phase by an integral multiple of 2π does not count in the values of the displacement and velocity.

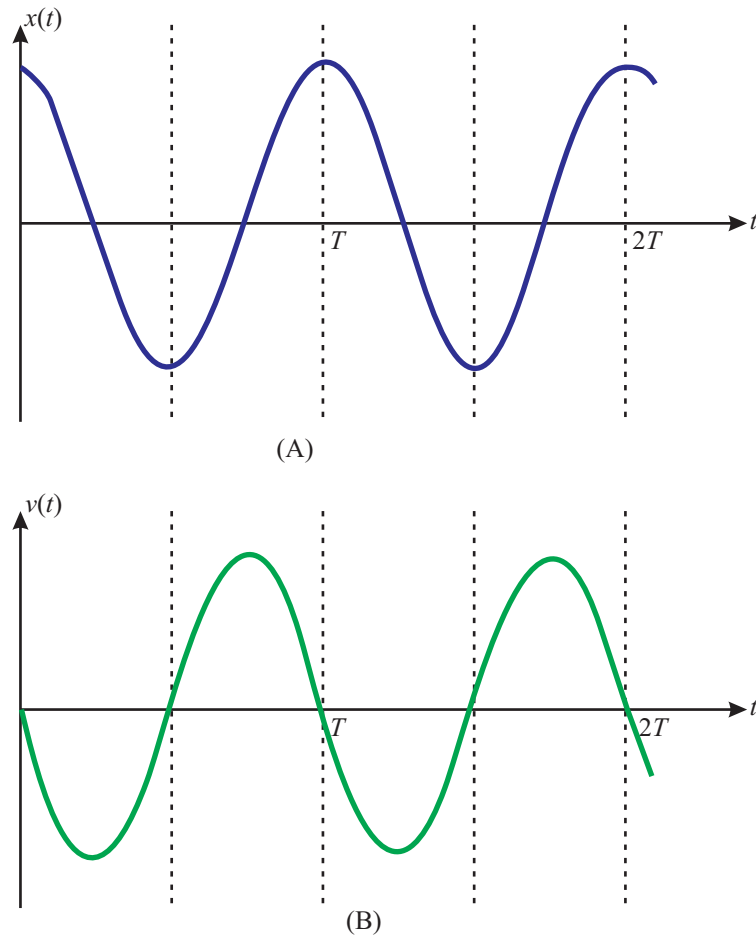


Figure 4-3: The variation of (A) displacement $x(t)$ with t , and (B) velocity $v(t)$ with t in simple harmonic motion; the initial conditions have been chosen as $x(t = 0) = x_0(> 0)$, and $v(t = 0) = 0$, which means that the amplitude is $a = x_0$.

4.2 Energy in simple harmonic motion

4.2.1 Potential energy

The expression for the potential energy of a particle at any point in a field of force involves an additive constant depending on the choice of a reference point. In the case of a simple harmonic motion this reference point is commonly taken to be the mean position of the particle. Choosing the line of motion of the particle as the x -axis and the mean position as the origin, the force on the particle is given by the expression

$$F(x) = -kx. \quad (4-22)$$

Eq. (3-59) then implies that the potential energy of the particle at the point x is

$$V(x) = \int_0^x kx \, dx = \frac{1}{2}kx^2, \quad (4-23a)$$

where the reference point is taken to be $x = 0$, the mean position of the particle.

In the above expression the symbol x has been used in two different senses - once as the upper limit of integration and again as the variable of integration. The variable of integration is a *dummy* variable, and can be represented by any other appropriate symbol as well. The way it has been written here is, however, a commonly accepted one.

Making use of eq. (4-2), one obtains an alternative expression for the potential energy:

$$V(x) = \frac{1}{2}m\omega^2 x^2. \quad (4-23b)$$

The variation of the potential energy with the position of the particle is shown graphically in fig. 4-4. Notice that the potential energy is zero at the mean position ($x = 0$; this is because the mean position has been chosen to be the reference point), and it increases along a parabolic curve on either side of the mean position .

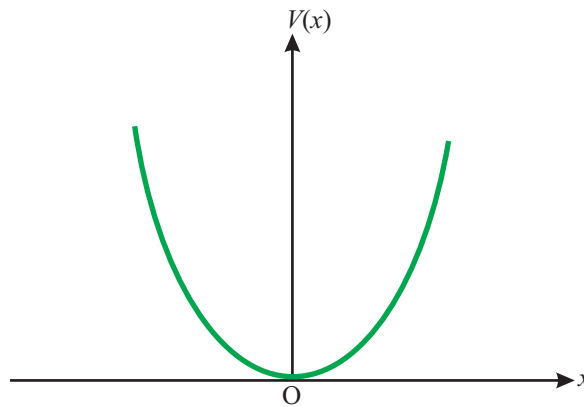


Figure 4-4: depicting schematically the variation of potential energy in a simple harmonic motion with distance from the mean position.

The fact that the potential energy is a minimum at $x = 0$ corresponds to the point $x = 0$ being the equilibrium position for the particle under consideration (and is independent of the choice of the reference point in arriving at (4-23b)): if the particle is located at any instant at $x = 0$ and its instantaneous velocity v is also zero, then it continues to be at the same point at all instants of time (refer to eq. (3-70)).

4.2.2 Kinetic energy in simple harmonic motion

The kinetic energy (K) of a particle was defined in section 3.2.2.7 as

$$K = \frac{1}{2}mv^2, \quad (4-24a)$$

where we saw that the difference in the kinetic energies at any two positions (say, P and Q) of the particle is the work done by the force acting on it as it moves from Q to P:

$$K_P - K_Q = \int_Q^P F(x)dx, \quad (4-24b)$$

where the integration is to be performed over the interval from Q to P.

Looking at eq. (4-15), the expression for the kinetic energy at the position x for a particle in SHM is seen to be

$$K(x) = \frac{1}{2}m\omega^2(a^2 - x^2). \quad (4-24c)$$

Equations (4-23b) and (4-24c) tell us that the *sum* of the potential and kinetic energies of the particle at the position x is given by

$$E = K(x) + V(x) = \frac{1}{2}mv^2 + \frac{1}{2}m\omega^2x^2 = \frac{1}{2}m\omega^2a^2. \quad (4-25)$$

This shows that, while the kinetic and the potential energies, considered separately, depend on the instantaneous position of the particle, their sum, the *total* energy (E) does not: the energy is conserved in a simple harmonic motion with a given angular frequency and amplitude. This is an instance of the principle of conservation of energy

for a particle in one dimensional motion.

Note that eq. (4-25) is nothing but eq. (4-4) in a different form (check this out; express the constant of integration C_1 in eq. (4-4) in terms of the energy E). In other words, the energy balance equation implied by the principle of conservation of energy gives the first integral of the equation of motion, i.e., the result of performing an integration on the equation of motion that leads to a first order differential equation from a second order one.

This, incidentally, is a general principle of mechanics: the equation expressing the conservation of energy in a conservative force field gives one the first integral of the equation of motion. One then has to perform one remaining integration so as to arrive at the solution of the equation of motion in terms of the initial conditions, or of two constants related to the initial conditions. This is precisely what we did in arriving at eq. (4-8) from eq. (4-4).

Problem 4-3

Average kinetic and potential energies in SHM.

A particle of mass $m = 0.01$ kg executes a simple harmonic motion with angular frequency $\omega = 10.0$ s⁻¹, amplitude $a = 0.2$ m, and initial phase $\delta = 0$. Find its average potential energy and kinetic energy.

Answer to Problem 4-3

HINT: For a periodic function $f(t)$ with period T , the *average* value is given by $\bar{f} = \frac{1}{T} \int_0^T f(t) dt$.

Thus, average kinetic energy is $\bar{K} = \frac{1}{T} \int_0^T \frac{1}{2} m v(t)^2 dt$ and average potential energy is $\bar{V} = \frac{1}{T} \int_0^T \frac{1}{2} m \omega^2 x(t)^2 dt$.

Now use $x(t) = a \cos(\omega t + \delta)$, and $v(t) = -a\omega \sin(\omega t + \delta)$, and $T = \frac{2\pi}{\omega}$ to obtain $\bar{K} = \bar{V} = \frac{1}{4} m \omega^2 a^2$.

Substituting the given values of m, ω , and a , one obtains $\bar{K} = \bar{V} = 0.01$ J. the initial phase δ has no role to play in determining these averages.

4.3 Simple harmonic oscillations of physical quantities

In this chapter I have talked of simple harmonic motion of a particle in one dimension, where the displacement of the particle from its mean position and its instantaneous velocity are given by expressions of the form (4-7), (4-14). Such a variation with time is referred to as a *sinusoidal* one. The graph representing a sinusoidal variation looks as in fig. 4-3(A), (B).

Apart from the displacement and velocity of a particle in simple harmonic motion, one encounters physical quantities in various other contexts in physics where the variations of these quantities with time are of a similar sinusoidal nature. If $A(t)$ denotes the instantaneous value of such a quantity, then its variation will be of the form

$$A(t) = a \cos(\omega t + \delta), \quad (4-26)$$

where a , ω , and δ are constants characterizing the variation of the quantity under consideration. As in the case of a simple harmonic motion of a particle, a represents the *amplitude* of variation of the physical quantity under consideration (which means that it varies periodically from $-a$ to a ; a is commonly chosen to be positive), while ω stands for the angular frequency, being related to the frequency (ν) and time period (T) of variation as $\omega = 2\pi\nu$ and $\omega = \frac{2\pi}{T}$ respectively.

These physical quantities are then said to undergo *simple harmonic oscillations*. In any such oscillation, the constants a , ω , and δ are referred to as the amplitude, angular frequency, and initial phase. The quantity $\Phi = \omega t + \delta$ is termed the *phase*. The value of this quantity modulo 2π , which we denote by ϕ , is also commonly referred to as the phase. For instance, the value $\Phi = \frac{7\pi}{2}$ corresponds to $\phi = \frac{3\pi}{2}$. The constant δ in the expression (4-26) stands for the constant part of the phase, or the *initial phase*.

The term oscillation is used in general to refer to alternating increase and decrease in the value of a quantity, which may not be periodic in the strict sense of the term. For instance, the damped vibration of a particle or a body is termed an oscillation in this

general sense. On the other hand, the term oscillation is also used in a more specific sense to refer to a periodic variation. The simplest instance of a periodic variation corresponds to a simple harmonic oscillation.

The velocity and acceleration of a particle executing a simple harmonic motion under a restoring force proportional to its instantaneous displacement, are instances of physical quantities undergoing simple harmonic oscillation. In a different context, one can think of the pressure at a point in a medium through which a sound wave of some particular frequency is propagating (see chapter 9 for necessary background). This pressure varies about a mean value in a sinusoidal manner, providing another instance of simple harmonic motion of a physical quantity. In a similar manner, the variation of the electric or magnetic field strength at any point in a medium through which an electromagnetic wave of a specific frequency is propagating (see chapter 14) is also simple harmonic in nature.

Yet another instance of simple harmonic oscillation is provided by the variation of current in an AC electrical circuit consisting of a capacitance and an inductance (see sec 13.5.2).

In all these instances, the variation of the relevant physical quantity (we denote this by, say, $u(t)$; at times the dependence on time t may be left implied) is described by some differential equation or other characterizing the variation, where the differential equation may be of the form

$$\frac{d^2u}{dt^2} = -\omega^2 u. \quad (4-27)$$

This is essentially the same as eq. (4-1), (4-2), though $u(t)$ may stand for a physical quantity different from $x(t)$.

However, this *need not* always be the differential equation in terms of which the time variation of u is determined in the first place. For instance, the differential equation determining the current as a function of time in an AC circuit looks like the equation of a

forced simple harmonic oscillator. What eq. (4-27) tells us is that $u(t)$ varies sinusoidally with time. Thus, the term ‘simple harmonic’ may have two connotations, depending on the context: *first*, a time variation where the defining equation is of the form (4-27), such as the equation for the displacement of a particle executing SHM along a straight line, and *secondly*, one in which the defining differential equation may differ, but the resulting time variation is a sinusoidal one.

Fig. 4-5(A) and (B) depict periodic but *non-sinusoidal* variations with time of a physical quantity u . One observes that u varies periodically with a period T , but the nature of the variation is otherwise different compared to that in fig. 4-3 (A), (B). Periodic variations of physical quantities in real life situations are, in general, non-sinusoidal. However, a non-sinusoidal periodic variation can be expressed as superpositions of sinusoidal components, where the frequencies of these sinusoidal components are multiples of a certain basic frequency characterizing the periodic variation under consideration.

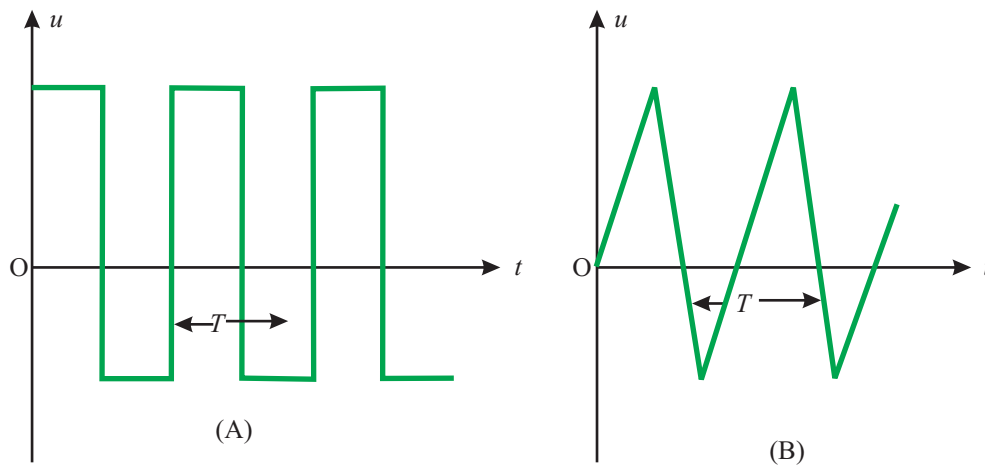


Figure 4-5: (A) and (B): Instances of non-sinusoidal periodic variations of a physical quantity u with time; T denotes the time period of the variation.

4.3.1 Angular oscillations

Imagine a rigid body rotating about an axis AB under the action of a torque, where the torque about the axis is proportional to the angular displacement from a mean position of the body, and the sense of rotation associated with the torque is always

directed towards the mean position. If M be the torque about the axis AB acting on the body at any instant of time t (where the sense of rotation associated with the torque is determined by the sign of M) and θ be the angular displacement then the angular motion is described by the equation (refer to eq. (3-163))

$$I \frac{d^2\theta}{dt^2} = M = -k\theta, \quad (4-28)$$

where I stands for the moment of inertia of the body about the axis AB, and k denotes the torque per unit angular displacement. This equation is of the form (4-27) and implies that the angular displacement θ executes simple harmonic oscillations under the action of the restoring torque M . The angular frequency of the oscillations is given by

$$\omega = \sqrt{\frac{k}{I}}. \quad (4-29)$$

As an example of angular oscillations, consider a rigid body freely suspended from a point A and capable of rotational motion about a horizontal axis (say, AB) through A. If C be the center of mass of the body (which rotates in a plane passing through A and perpendicular to the axis AB), then the forces acting on the body are its weight W acting vertically downward through C, and the reaction at the support at A.

The total moment of these forces about the axis AB is (see fig. 4-6) $M = -Wl \sin \theta$ where l stands for the distance AC, and θ denotes the instantaneous value of the angle between AC and the vertical direction (measured in the sense of the bent arrow in the figure). The negative sign in the expression for the torque M signifies that the torque is restoring in nature, i.e., the sense of rotation associated with it is such as to bring the line AC to the vertical position.

Assuming that the angular displacement θ is sufficiently small, one can approximate the torque as $M = -Wl\theta$, and the rotational motion of the rigid body (referred to as a *compound pendulum*) is then described by eq. (4-28), where $k = Wl$. The angular

frequency of the simple harmonic oscillation of θ is thus

$$\omega = \sqrt{\frac{Wl}{I}}, \quad (4-30a)$$

while the corresponding time period is

$$T = \frac{2\pi}{\omega} = 2\pi\sqrt{\frac{I}{Wl}}. \quad (4-30b)$$

At times, it is convenient to express the frequency or the time period of angular oscillations in terms of the radius of gyration (see section 3.19.13.3) of the body about the axis (AB) through the point of suspension, or of the radius of gyration about a parallel axis passing through the center of mass of the body. Denoting the two radii of gyration by K and K_{CM} respectively, one obtains

$$\omega = \sqrt{\frac{gl}{K^2}} = \sqrt{\frac{gl}{K_{CM}^2 + l^2}}, \quad (4-31a)$$

$$T = 2\pi\sqrt{\frac{K^2}{gl}} = 2\pi\sqrt{\frac{l + \frac{K_{CM}^2}{l}}{g}}, \quad (4-31b)$$

where the relation (3-173) has been made use of (replacing d with l in keeping with the present notation). It may be mentioned that these formulae for the compound pendulum resemble the corresponding expressions for a *simple* pendulum (equations (4-37a), (4-37b) in sec. 4.4.1) provided that the length l of the latter is replaced with the *effective* length $l_{\text{eff}} = l + \frac{K_{CM}^2}{l}$ of the former, where the symbol l on the right hand side stands for the distance of the center of mass of the compound pendulum from the axis of oscillation.

Problem 4-4

A homogeneous lamina of square shape is suspended from one corner so that it can swing freely in a vertical plane. If the mass of the lamina is M and the length of a side of the square is a , find the time period of angular oscillations of the lamina with a small angular amplitude .

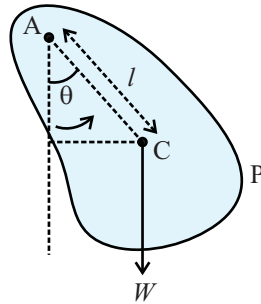


Figure 4-6: Illustrating the angular oscillations of a compound pendulum; a rigid body P is suspended from the point A such that it can rotate freely about a horizontal axis AB (not shown in the figure) through A; C is the center of mass of the body through which its weight W acts vertically downwards; the angular displacement of AC from the vertical line is θ measured in the sense of the bent arrow; for small values of θ , the torque acting on the body is $-Wl\theta$ and is of a restoring nature, tending to rotate AC towards the vertical position; the body executes angular oscillations of the simple harmonic type about the axis AB.

Answer to Problem 4-4

HINT: The moment of inertia of the lamina about an axis passing through one corner and perpendicular to its plane (the axis of rotation in the present context) is $I = \frac{2}{3}Ma^2$ (check this out by working out the moment of inertia about an axis along one edge, which turns out to be $I' = \frac{1}{3}Ma^2$, and then applying the principle of perpendicular axes, which gives $I = 2I'$). The distance of the center of mass from the center of suspension is $l = \frac{a}{\sqrt{2}}$. Now apply eq. (4-30a) to obtain $\omega^2 = \frac{3\sqrt{2}}{4} \frac{g}{a}$. Finally, $T = \frac{2\pi}{\omega}$.

Problem 4-5

A rigid frame made up of a light horizontal platform AB and two light rods SA and SB is free to rotate about a horizontal axis through a point of suspension S (fig. 4-7). A little mouse moves on the platform in such a manner that the platform remains at rest. Show that the mouse executes a simple harmonic motion. Find an expression for the time period of the motion and also for the maximum length of the platform for which such a motion is possible. Assume that the coefficient of limiting static friction between the mouse and the platform is μ , and that the mouse spans the entire length (l) of the platform in its motion.

Answer to Problem 4-5

HINT: Let P be the instantaneous position of the mouse (which we assume to be a particle) at

any given time t , where SP makes an angle θ with the vertical line, as shown in fig. 4-7. In order that the platform may remain at rest, the turning moment of the resultant force on the platform has to be zero, i.e., the line of action of the resultant has to pass through S. The force of friction exerted by the mouse on the platform is, say, F at the chosen instant, acting in the direction shown, while the weight of the mouse of mass, say, m is mg acting vertically downward on the platform (g =acceleration due to gravity). One thus has to have $\tan \theta = \frac{x}{h} = \frac{F}{mg}$, where x stands for the instantaneous displacement with reference to the mean position O, and h for the distance SO (reason out why; see fig. 4-7).

The mouse moves by pushing against the platform in such a manner that, at any instant of time, some of its feet are at rest on it while the others are raised for forward stepping, and it has to adjust the force of static friction such that the above equality is satisfied. The forward force on the mouse is $F = \frac{mgx}{h}$ acting towards O. This being proportional to the displacement x , the resulting motion is a simple harmonic one, with a force constant $k = \frac{mg}{h}$, and hence a time period $T = 2\pi\sqrt{\frac{h}{g}}$. For such a motion to be possible, with the mouse spanning the entire length (l) of the platform, one must have $F|_{x=\frac{l}{2}} \leq \mu mg$, i.e. $\frac{l}{2h} \leq \mu$ (reason out why).

Problem 4-6

Referring to fig. 4-6 write down the differential equation for angular oscillations of a compound pendulum *without assuming the angular amplitude (θ_0) to be small*; work out an expression for the force of reaction on the compound pendulum at the point of suspension A when the angular displacement from the mean position is θ .

Answer to Problem 4-6

HINT: The forces acting on the rigid body making up the compound pendulum are the reaction at the point of suspension A, and the resultant gravitational force mg passing through the center of mass C, acting in the vertically downward direction. Let the components of the reaction force, resolved along CA be N_1 and that along a perpendicular direction (see fig. 4-8) be N_2 . The torque due to the gravitational force about the point A is $mg l \sin \theta$ which constitutes the total moment of the forces about A (the moment of the reaction force is zero since it acts through A). Referring to formula (3-163) and making use of the notation employed in the present section, the differential

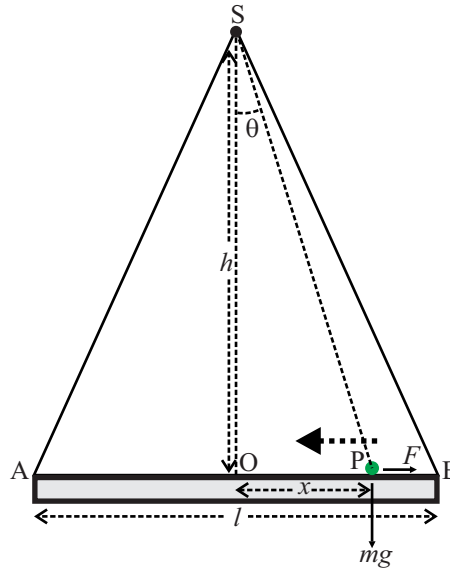


Figure 4-7: A mouse running to-and-fro on a light platform AB, suspended with light rods SA, SB from a point S such that the entire frame can rotate freely about S as a rigid system; the wise mouse makes use of the force of friction F in such a manner that the frame remains at rest (problem 4-5); P denotes the position of the mouse at any chosen instant t , the instantaneous direction of motion being shown by the dotted arrow; the resultant of the force F exerted by the mouse on the platform, and the weight mg acting on the platform has to pass through S; this requires $F = \frac{mgx}{h}$, where x, h are distances shown; for the motion to be possible with the mouse spanning the entire length of the platform, one has to have $\mu > \frac{2h}{l}$, where μ is the coefficient of static friction.

equation describing the angular oscillations is seen to be

$$I \frac{d^2\theta}{dt^2} = -mgl \sin \theta, \quad (4-32)$$

(check this out). On multiplying both sides by $2 \frac{d\theta}{dt}$ and performing an integration, one obtains

$$I \left(\frac{d\theta}{dt} \right)^2 = 2mgl (\cos \theta - \cos \theta_0), \quad (4-33)$$

(check *this* out as well; you will have to work out a few steps of calculus). Now consider the center of mass motion of the body (i.e., to the equivalent motion of a particle of mass m , imagined to be placed at the center of mass, moving under the joint action of the weight of the body and the reaction force (with components N_1, N_2), imagined to be acting through C). In accordance with the results of sections 3.17.4 and 3.19.5, the radial and cross-radial accelerations describing this

motion are given by (refer to the second relation in (3-60))

$$ml\left(\frac{d\theta}{dt}\right)^2 = -mg \cos \theta + N_1, \quad (4-34a)$$

$$ml\frac{d^2\theta}{dt^2} = -mg \sin \theta - N_2, \quad (4-34b)$$

(check this out; note that we are now considering the motion of the center of mass, and not that of the body as a whole). Making use of formulae (4-32), (4-33), one obtains the components of the reaction force as

$$N_1 = mg\left(\cos \theta + 2\frac{l^2}{K^2}(\cos \theta - \cos \theta_0)\right), \quad N_2 = -mg \sin \theta\left(1 - \frac{l^2}{K^2}\right). \quad (4-35)$$

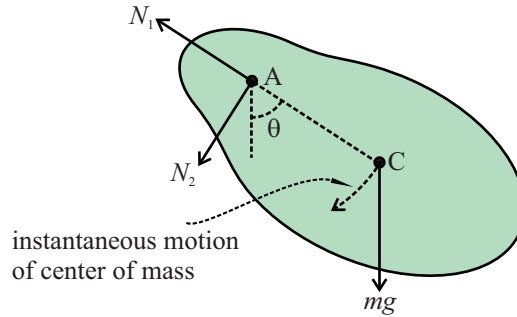


Figure 4-8: Illustrating the components (N_1, N_2) of the reaction force at the point of suspension of a compound pendulum at an instant when the instantaneous angular displacement from the mean position is θ (refer to problem 4-6; see also fig. 4-6); the angular amplitude θ_0 need not be small; N_1, N_2 are determined by referring to the center of mass motion, i.e., the motion of a particle of mass m (the mass of the rigid body) imagined to be placed at the center of mass C , under the action of the gravitational force mg and the forces N_1, N_2 , all imagined to be acting at C .

4.4 The pendulum and the spring

4.4.1 The simple pendulum

The mechanical system described in section 4.1, is referred to as a *plane pendulum*. It is made up of a light rigid rod fixed at its upper end, with a small point-like mass (the 'bob')

attached to its lower end. When the rod oscillates in a vertical plane about its mean, vertical, position (the line SO in fig. 4-1), the bob performs a to-and-fro motion along a circular arc centered about the fixed upper end of the rod. The two extreme positions of the bob on the two sides of the mean position O in fig 4-1 are A and A'. If the distance of either of these from the mean position is small then in such a special situation the bob (which can be taken to be a point mass in the present context) can be assumed, in an approximate sense, to move on a straight line, on which it executes a simple harmonic motion. A plane pendulum oscillating with a small amplitude is commonly referred to as a *simple pendulum*, the oscillations of the latter being simple harmonic in nature.

With reference to the motion of a simple pendulum, a quantity of greater relevance compared to the displacement of the bob along a straight line, is the *angular displacement* (say, θ) of the pendulum rod from the vertical position SO (see fig. 4-9).

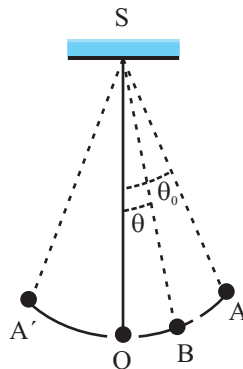


Figure 4-9: Angular amplitude and angular displacement of a simple pendulum; SO denotes the vertical position while SA and SA' denote the positions with maximum magnitude of angular displacement (θ_0) on either side of the mean position; SB denotes an arbitrarily chosen position with angular displacement θ ; the variation of θ with time constitutes a simple harmonic oscillation if θ_0 is sufficiently small.

Notice that the *same* motion can be looked upon from two points of view: a circular motion of the rigid body made up of the rod and the bob about a horizontal axis passing through the point of support S and perpendicular to the plane of the diagram in fig 4-9, and at the same time, a circular motion of the point-like bob about the point S (which, in an approximate sense, is a linear motion, provided that the angular

amplitude of the pendulum is small), where the motion of the rod is ignored because of its negligible mass. The use of the angular displacement θ is convenient in both descriptions.

As mentioned in sec. 3.19.7.1 (see eq. (3-149)), the rate of change of angular momentum of a particle about any given point is equal to the moment of the force acting on the particle. One can make use of this principle in deriving the equation that describes the circular motion of the bob of the simple pendulum, which is a special case of the motion considered in sec. 4.3.1. One finds this equation to be

$$l \frac{d^2\theta}{dt^2} = -g \sin \theta, \quad (4-36a)$$

where l denotes the length of the pendulum rod and g the acceleration due to gravity (check this out).

If now the angular amplitude θ_0 of the oscillation is assumed to be small then one can employ the approximation $\sin \theta \approx \theta$ and arrive at a simpler form of the above equation:

$$\frac{d^2\theta}{dt^2} = -\frac{g}{l} \theta. \quad (4-36b)$$

This is seen to be of the same form as eq. (4-1) or (4-27), which means that the variation of the angular displacement of a simple pendulum is a simple harmonic oscillation if the angular amplitude happens to be sufficiently small. The angular frequency of the oscillation is given by

$$\omega = \sqrt{\frac{g}{l}}, \quad (4-37a)$$

while the expression for the time period is (see eq. (4-18))

$$T = 2\pi \sqrt{\frac{l}{g}}. \quad (4-37b)$$

However, even for a small angular amplitude, there still remains an error in the above

approximation and consequently in the formula (4-37a), (4-37b) for the angular frequency and the time period. The error diminishes monotonically with the amplitude.

Strictly speaking, the equation of motion of the system under consideration is eq. (4-36a), and not eq. (4-36b), which means that while its motion is periodic, it is not in reality a simple harmonic one. If the angular amplitude of the motion is θ_0 , then a more accurate expression for the time period of the motion works out to

$$T = 2\pi \sqrt{\frac{l}{g}} \left(1 + \frac{\theta_0^2}{16}\right). \quad (4-38)$$

Evidently, for small values of θ_0 , the formula (4-38) is closely approximated by (4-37b). Notice that, according to eq. (4-38), the time period T should depend on the angular amplitude θ_0 . However, for *small* values of the amplitude the approximate formula (4-37b) shows that the *time period is independent of the amplitude*. This is referred to as the *law of isochronicity* of the simple pendulum. Another principle of considerable relevance in this context is also implied by eq. (4-37b): *the time period of a simple pendulum is independent of the mass of the bob*. Notice that this latter principle remains valid even when the amplitude of motion is not small.

For sufficiently small amplitude of oscillation, the motion of the bob along the arc of a circle can be approximated as a motion along a straight line, and then the displacement x from the mean position may be seen to satisfy the differential equation (4-1), with $k = m \frac{g}{l}$.

The plane pendulum and the simple pendulum are idealized systems, constituting special cases of the more realistic *compound pendulum* referred to in sec. 4.3.1.

4.4.2 The spring

Another practical device that produces a simple harmonic motion is a light *spring* fixed at one end, with a particle (in practice a rigid body of small size) attached to the other end. The spring has a natural length (say, l), to which there corresponds an equilibrium

position of the particle where the spring does not exert any force on it. If now the spring is extended or compressed by a distance x along its length, then it tends to get back to its natural length due to its elastic property and exerts a force on the particle directed towards its equilibrium position. Assuming that the extension or contraction of the spring is sufficiently small, the restoring force can be expressed in the form $-kx$ where k is a constant (commonly referred to as the spring constant) determined by the elastic properties of the spring and x is commonly measured along the direction in which the spring gets extended (we assume that the extension and contraction of the spring takes place along a straight line which can be taken as the x-axis).

The equation of motion of the particle is then seen to be the same as eq. (4-1), where m once again stands for the mass of the particle. Thus, if the equilibrium of the system made up of the spring and the particle is disturbed by extending or compressing the spring and then letting it go or, say, by giving it a blow, the particle will execute a simple harmonic motion with an angular frequency $\omega = \sqrt{\frac{k}{m}}$.

Considering a configuration of the system corresponding to an extension x of the spring (a contraction can be considered as an extension with a negative sign), the potential energy of the particle is given by the expression (4-23a), which can also be interpreted as the energy of deformation of the spring. Since the spring is assumed to be a light one (which implies that its mass is vanishingly small), the kinetic energy of the spring can be taken to be zero. The total energy of the system, made up of the kinetic energy of the particle and its potential energy (which can be interpreted as the potential energy of deformation of the spring) is then given by the expression (4-25), which remains constant throughout the motion.

Problem 4-7

Consider a block attached to two springs, one end of each spring being attached to one of two fixed walls as shown in fig. 4-10. If the block moves on smooth horizontal ground find its frequency of oscillation, given that its frequencies when connected to only one of the two springs in succession are $f_1 = 1.0$ Hz and $f_2 = 1.5$ Hz.

Answer to Problem 4-7

HINT: The spring constants k_1 and k_2 are obtained from the relations $f_i = \frac{1}{2\pi} \sqrt{\frac{k_i}{m}}$ ($i = 1, 2$), where m stands for the mass of the block. If the displacement of the block from its mean position be x , where the positive direction of the x-axis is towards the right in the figure, the restoring force exerted on the block by the springs are respectively $-k_1x$ and $-k_2x$ (note that when one spring is extended, the other is compressed, and the force exerted by either of the springs always acts towards the mean position). Thus, the required frequency is $f = \frac{1}{2\pi} \sqrt{\frac{k_1+k_2}{m}} = \sqrt{(f_1^2 + f_2^2)}$. This gives, in the present instance, $f = 1.80$ Hz (approx).

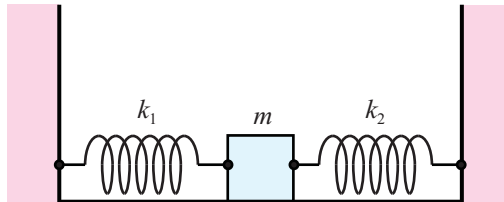


Figure 4-10: A block connected to two springs, where one end of each spring is attached to one of two fixed walls; the block can move on a smooth horizontal surface; when set in oscillation, the block performs a simple harmonic motion; during the time when one of the two springs is extended, the other is compressed, and the force exerted by either of the springs always acts towards the mean position of the block.

Problem 4-8

Consider two blocks of mass $m_1 = 5.0$ kg, and $m_2 = 2.0$ kg, the latter on top of the former, connected to one end a spring of spring constant $k = 50 \text{ N}\cdot\text{m}^{-1}$, as shown in fig. 4-11 whose other end is fixed to a wall. The combination of the blocks can move on a frictionless horizontal surface. Supposing that the system is initially at rest and then set in oscillation where both the blocks move together, and that the coefficient of friction between the blocks is $\mu = 0.20$, what is the maximum possible amplitude of oscillation? ($g = 9.8 \text{ m}\cdot\text{s}^{-1}$.)

Answer to Problem 4-8

The restoring force on the two blocks taken together is of maximum magnitude (F_0) when the displacement is $x = \pm a$, where a is the amplitude, and thus F_0 is given by $F_0 = ka$. This implies that, if the blocks move together without any relative motion between them, the maximum magnitude

of the force on the upper block (of mass m_2) is $F' = \frac{m_2 k a}{m_1 + m_2}$. This must be less than or equal to the maximum value of the frictional force on the upper block, i.e., the condition for the two blocks to move together is $F' \leq \mu m_2 g$. In other words, the maximum allowed value of the amplitude a is obtained from the relation $\frac{m_2 k a}{m_1 + m_2} = \mu m_2 g$. If the amplitude of oscillation exceeds this value, the upper block will slip over the lower. In other words, the required maximum value of the amplitude is $\frac{(m_1 + m_2) \mu g}{k} = 0.274 \text{ m (approx)}$.

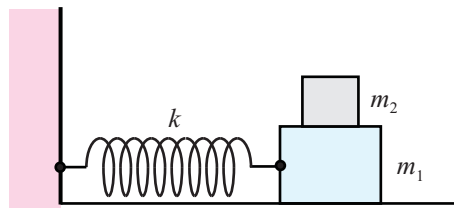


Figure 4-11: Two blocks, one on top of the other, attached to one end of a spring, the other end of which is attached to a fixed wall; the blocks can move on a smooth horizontal surface; if the system is made to oscillate with an amplitude a , then the blocks will move together up to a certain maximum value of a , above which the upper block will slide over the lower one.

Problem 4-9

A small bob of mass $m = 1 \text{ kg}$ is suspended from the lower end of a light spring whose upper end is attached to a fixed point. The body is displaced by $a = 0.1 \text{ m}$ below its position of equilibrium and then released from rest. If the spring constant is $k = 50 \text{ N} \cdot \text{m}^{-1}$, find the frequency of oscillation of the bob. What is the net force on the bob when its displacement from the position of equilibrium is $x = 0.05 \text{ m}$ in the vertically downward direction? For the spring-bob system, find the gravitational potential energy in this position, the deformation energy of the spring, and the total energy, with reference to the equilibrium configuration ($g = 9.8 \text{ m} \cdot \text{s}^{-2}$).

Answer to Problem 4-9

The equilibrium position of the bob corresponds to an extension of the spring from its natural length by an amount x_0 given by $kx_0 = mg$ (g = acceleration due to gravity). For a downward displacement x from the equilibrium position, the net force on the bob is $F = kx$ in the upward direction, i.e., towards the equilibrium position ($k(x + x_0)$ upward due to the spring and $mg = kx_0$ downward due to the pull of gravity). Using given values, this works out to $F = 2.5 \text{ N}$. The force

per unit displacement is k , and hence the frequency of oscillation is $f = \frac{1}{2\pi} \sqrt{\frac{k}{m}} = 1.125 \text{ Hz}$.

The gravitational potential energy (with reference to the equilibrium position) of the system for a displacement x is $E_1 = -mgx = -0.49 \text{ J}$ (the spring being a light one, its gravitational potential energy can be ignored). The deformation energy of the spring with reference to the equilibrium configuration is $E_2 = \frac{1}{2}k((x_0 + x)^2 - x_0^2) = \frac{1}{2}kx^2 + kxx_0 = \frac{1}{2}kx^2 + mgx = 0.553 \text{ J}$. Finally, the kinetic energy of the bob is (the spring being a light one its kinetic energy is negligible) $E_3 = \frac{1}{2}mv^2 = \frac{1}{2}k(a^2 - x^2)$, where $a = 0.1 \text{ m}$ stands for the amplitude. In other words, $E_3 = 0.185 \text{ J}$. Summing up, the total energy of the system, which remains constant during the oscillatory motion, is $\frac{1}{2}ka^2 = 0.25 \text{ J}$.

4.5 Damped simple harmonic motion

4.5.1 Damped SHM: equation of motion

We have seen that a restoring force acting on a particle causes the latter to execute a simple harmonic motion where the restoring force, acting toward a fixed point on the line of motion, is proportional to the instantaneous displacement from that fixed point. In addition, a *damping* force may act on the particle whose effect is to *retard* its motion, i.e., to decrease the magnitude of its instantaneous velocity. Such retarding forces may arise, for instance, due to the viscous resistance to motion offered by the medium in which the motion occurs or due to various other factors related to friction.

The direction of a retarding force is opposite to the instantaneous velocity of the particle, and its magnitude may be assumed to be proportional to the magnitude of the velocity, provided the latter lies within certain limits. The expression for the retarding force is therefore of the form $-\gamma v$, where γ is a positive constant, referred to as the *retardation constant*.

Assuming a restoring force and a retarding force to act on a particle *simultaneously*, one

can write the equation of motion in the form

$$m \frac{d^2 x}{dt^2} = -kx - \gamma v. \quad (4-39)$$

This equation is obtained by adding the term $-\gamma v$, the retarding force acting in addition to the restoring force, to the right hand side of eq. (4-1), where the latter describes a simple harmonic motion caused by the restoring force alone.

1. While the system I am talking of is a particle in motion, one can also talk of damped oscillations of a body of finite size where the latter executes a repetitive motion experiencing, at the same time, a resistance to the motion. For instance, when a tuning fork is set in vibration, any one arm of the tuning fork executes a damped vibration - it vibrates on either side of its mean position while the amplitude of vibration goes on decreasing. We shall see that this is the general nature of the solution to eq. (4-39). What is common between the vibrating tuning fork and the particle referred to above is the equation governing their motion, which is of the general form (4-39).
2. We refer here to a one dimensional motion of a particle along a straight line, which we take as the x-axis though, as mentioned above, motions of various other descriptions may also be described in similar terms. Quantities like the displacement x and the velocity v are then to be interpreted as the components of the corresponding vectors along the x-axis. These carry their appropriate signs and represent one-dimensional vectors. Any of the two possible directions along the line of motion can be chosen as the positive direction of the x-axis, the common choice being the one from the left to the right.
3. The restoring force and the retarding force on the particle have been assumed to be proportional to the displacement and velocity respectively, though there is no fundamental reason why this should be so. The dependence of the force on displacement and velocity may more generally be of a different type. In numerous situations of interest, however, the force depends *linearly* on the displacement and velocity as in eq. (4-39) for sufficiently small magnitudes of these two quantities.

Eq. (4-39) is a second order differential equation like eq. (4-1), and one can determine the

general solution to this equation in a manner similar to that for the latter. The general solution once again involves two undetermined constants. These can be determined by making use of the *initial conditions* of motion, i.e., the co-ordinate x_0 and the velocity v_0 of the particle at any given instant of time, say, $t = 0$. One thereby arrives at a *particular solution* of the differential equation describing the damped simple harmonic motion.

The nature of the general solution or of a particular solution depends on the values of the parameters m , k , and γ . It turns out that the use of the parameters

$$\omega = \sqrt{\frac{k}{m}}, \quad (4-40a)$$

and

$$b = \frac{\gamma}{2m}, \quad (4-40b)$$

is more convenient in describing the solutions. Of these, the former is the frequency of the oscillation *in the absence of the damping force*, while the latter is referred to as the *damping constant*. In reality, the nature of the motion resulting from eq. (4-39) depends on the *relative magnitude* of these two parameters. In terms of the parameters b and ω , the differential equation (4-39) reads

$$\frac{d^2x}{dt^2} + 2b\frac{dx}{dt} + \omega^2x = 0. \quad (4-41)$$

The role of the parameters b and ω in determining the nature of the motion can now be indicated. For instance, if b is less than ω ($\frac{b}{\omega} < 1$), i.e., the damping constant is relatively small, the motion is said to be an *underdamped* one while a relatively strong damping ($\frac{b}{\omega} > 1$) corresponds to what is termed an *overdamped* motion. Finally, the special case $\frac{b}{\omega} = 1$ is said to correspond to *critical damping*.

4.5.2 Underdamped and overdamped motions

4.5.2.1 Underdamped SHM

Fig. 4-12 depicts graphically the nature of variation of the displacement of the particle with time for an underdamped simple harmonic oscillation. Comparing this with fig. 4-3(A), one can have an idea of the effect of damping on a simple harmonic motion. In this graph, which has been drawn schematically for a particular initial condition, look at the points corresponding to maximum displacement on either side of the point $x = 0$ (the displacement x is plotted along the y-axis, with the time t plotted along the x-axis), i.e., the one toward which the restoring force acts. The value of the maximum displacement is seen to decrease in successive swings of the graph, there being alternating episodes of maximum displacement on the two sides of the mean position.

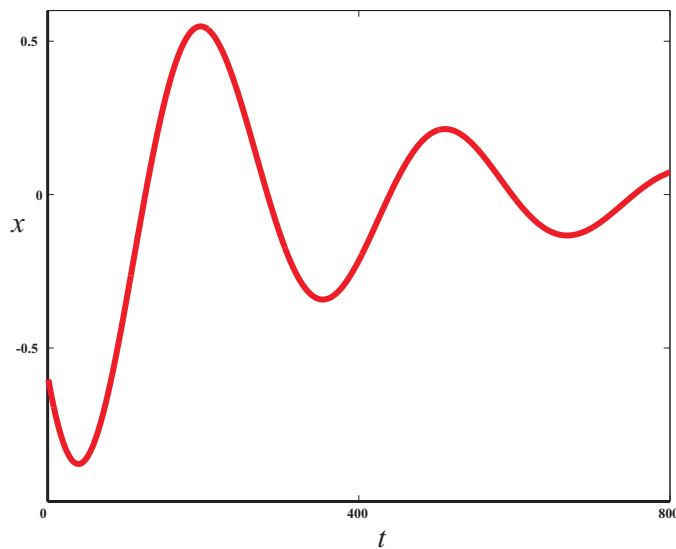


Figure 4-12: Variation of displacement ($x(t)$, plotted along the y-axis in the figure) with time (along the x-axis) in an underdamped simple harmonic motion; the particle oscillates on either side of the point $x = 0$, but the successive values of maximum displacement on any one side go on decreasing; parameters chosen arbitrarily; units along the x- and y-axes also chosen arbitrarily, indicating only the nature of variation.

Referring to the successive maximum displacements on either side of the point $x = 0$ as the successive ‘amplitudes’, one concludes that, in contrast to an undamped simple harmonic motion where the amplitudes are all equal, those in an underdamped simple

harmonic motion *go on decreasing*.

The expression for the displacement as a function of time in an underdamped simple harmonic motion is of the form (recall that the condition for underdamped motion is $b < \omega$)

$$x(t) = ae^{-bt} \cos(\omega' t + \delta), \quad (4-42a)$$

where a and δ are two constants depending on the initial conditions of motion, and ω' is given by

$$\omega' = \sqrt{\omega^2 - b^2} \quad (b < \omega). \quad (4-42b)$$

Expression (4-42a) may be compared with the corresponding expression in (4-7). While the angular frequency in the absence of damping is ω (and correspondingly, the frequency is $\nu = \frac{\omega}{2\pi}$), it is modified to ω' due to the effect of damping, where $\omega' < \omega$. In other words, the effect of damping is to reduce the frequency. However, if the damping constant b is small compared to ω (i.e., $\frac{b}{\omega} \ll 1$) this decrease in frequency may be ignored from a practical point of view.

The next interesting feature to note in eq. (4-42a) is that it involves the expression ae^{-bt} in place of a of eq. (4-7). We have seen that a represents the amplitude of an undamped simple harmonic motion, since it corresponds to the maximum magnitude of the displacement on either side of the mean position, for values ± 1 of the factor $\cos \omega t$ in eq. (4-7). In the presence of damping, however, the value of the expression ae^{-bt} gradually decreases at the successive instants when $\cos \omega t$ attains the values ± 1 . In other words, the successive amplitudes go on decreasing, as seen from fig. 4-12.

In summary, then, the nature of an underdamped simple harmonic motion is similar to that of an undamped one. In both, there occurs a to-and-fro motion of the particle about a fixed point on a straight line, though the frequency of oscillation differs in the two motions, and the magnitude of the maximum displacement on either side of the fixed point decreases in successive oscillations in the damped motion. Strictly speaking,

the damped motion is not a periodic one, but it is still referred to as an oscillation.

4.5.2.2 Overdamped SHM

The *overdamped* simple harmonic motion is of a different nature. In this case there does not occur an alternating increase and decrease in the displacement or the velocity of the particle. let us assume, without loss of generality, that the particle is initially located at a point in the positive half of the x-axis ($x(0) > 0$), where the restoring force acts towards the point $x = 0$. If the initial velocity is zero ($\dot{x}(0) = 0$) then the particle will monotonically approach the origin. If, on the other hand the initial velocity differs from zero then one of the following three alternatives may occur: (a) for $\dot{x}(0) > 0$ the particle moves away from the origin in the positive half of the x-axis and, after attaining its maximum displacement, returns monotonically to the origin; (b) for $\dot{x}(0) < 0$, but sufficiently small in magnitude, the particle once again approaches the origin monotonically; (c) for $\dot{x}(0) < 0$ and of sufficiently large magnitude, the particle crosses over to the negative half of the x-axis and, after attaining the maximum displacement, monotonically approaches the origin.

In other words, the motion can no longer be termed an oscillation or a vibration. A typical time-displacement graph representing the motion (corresponding to case (c) above) looks as in fig. 4-13. The solution to eq. (4-41) for the displacement as a function of time looks like (recall that the condition for overdamped motion is $b > \omega$)

$$x(t) = e^{-bt}(Ae^{\sqrt{(b^2-\omega^2)t}} + Be^{-\sqrt{(b^2-\omega^2)t}}) \quad (b > \omega), \quad (4-43)$$

where A, B that get determined if one knows the initial conditions.

In between the undamped and the overdamped simple harmonic motions, one encounters the *critically damped* simple harmonic motion in the special situation corresponding to $b = \omega$. Here again, the motion of the particle is not an oscillatory one, being somewhat

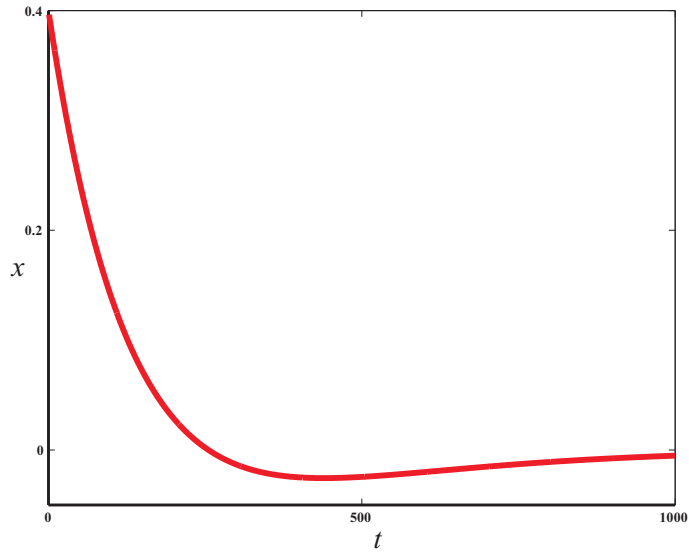


Figure 4-13: Variation of displacement with time in an overdamped simple harmonic motion; the particle does not oscillate on either side of the point $x = 0$, and approaches this point monotonically once the maximum displacement on any one side is reached; parameters chosen arbitrarily; units along the x- and y-axes (time and displacement respectively) are also chosen arbitrarily, indicating only the nature of variation.

similar in nature to the overdamped motion. The general solution of (4-41) is of the form

$$x(t) = e^{-bt}(\alpha + \beta t) \quad (b = \omega), \quad (4-44)$$

where α, β are constants, with values depending on the initial conditions (i.e., x, \dot{x} at $t = 0$).

Problem 4-10

Establish formulae (4-42a), (4-43), (4-44) for $b < \omega, b > \omega, b = \omega$ respectively.

Answer to Problem 4-10

Define a new variable $y(t)$ as $y(t) = e^{bt}x(t)$, and set up the equation for $y(t)$ by substitution in eq. (4-41).

Problem 4-11

Referring to an overdamped simple harmonic motion of a particle represented by eq. (4-41), suppose that the particle is initially at $x_0 < -a$ ($a > 0$) and has an initial velocity $\dot{x}(0) = u > 0$. Find the minimum value (u_0) of u such that the particle crosses over to the side $x > 0$ and, for $u > u_0$, the time at which the crossing takes place. What is time at which the particle attains maximum displacement from the origin on the side $x > 0$?

Answer to Problem 4-11

HINT: Let us, for the sake of simplicity, define $q_1 = b - \sqrt{b^2 - \omega^2}$, $q_2 = b + \sqrt{b^2 - \omega^2}$ (recall that, for overdamped motion, $b > \omega$). Then, referring to the formula (4-43), one obtains

$$x(t) = Ae^{-q_1 t} + Be^{-q_2 t}, \quad \dot{x}(t) = -(q_1 Ae^{-q_1 t} + q_2 Be^{-q_2 t}),$$

where $A = \frac{-aq_2 + u}{q_2 - q_1}$, $B = \frac{aq_1 - u}{q_2 - q_1}$. The condition for crossing is that $x(t) = 0$ for some positive value of t , which requires that $u > aq_2$ (reason this out). Hence the required minimum value of the initial velocity above which the particle crosses over to the positive half of the x-axis is $u_{\min} = a(b + \sqrt{b^2 - \omega^2})$. For $u > u_{\min}$, the time at which maximum displacement is attained after crossing corresponds to $\dot{x}(t) = 0$, i.e., $e^{(q_2 - q_1)t} = -\frac{Bq_2}{Aq_1}$.

Analogous to the simple harmonic variation of physical quantities in various physical contexts other than the motion of a particle, one encounters *damped* simple harmonic variations of various different physical quantities as well. For instance, the variation of current or voltage in an electrical circuit made up of an inductance (L), a capacitance (C), and a resistance (R) can be described in terms of an equation similar to (4-39). Here again, one can have underdamped, overdamped, or critically damped variation of the physical quantity under consideration, i.e., the values of the circuit elements (L, C, R).

4.5.3 Damped SHM: dissipation of energy

In the presence of damping, the energy of the oscillator does not remain constant during its motion. Consider a particle executing a damped simple harmonic motion described by eq. (4-39). Let the particle be at the point P with co-ordinate x_1 at time t_1 , and at Q with co-ordinate x_2 at a later time $t_2 (> t_1)$. Then the work done by the restoring force,

which we define as the decrease of potential energy (corresponding to the restoring force considered in isolation, without reference to the damping force) from P to Q is given by

$$W|_{P \rightarrow Q} = -(V_Q - V_P) = K_Q - K_P + \gamma \int_{t_1}^{t_2} v^2 dt, \quad (4-45)$$

where v denotes the velocity of the particle at time t , and the kinetic energy is defined in the usual manner (recall that the definition of the kinetic energy does not require any reference to the force).

Problem 4-12

Check eq. (4-45) out.

Answer to Problem 4-12

Integrate both sides of eq. (4-39) with respect to x from P to Q, and make use of the relations

$$\frac{dv}{dt} dx = v dv, \quad v dx = v^2 dt.$$

Since $\gamma > 0$, the relation (4-45) shows that, for $t_2 > t_1$, the work done on the particle, which can be interpreted as the energy given to it in its motion from P to Q, is *greater* than the increase in its kinetic energy by the amount $\int_{t_1}^{t_2} v^2 dt$, i.e., in other words, some of the work done does not appear as the kinetic energy of the particle. In reality, this part of the energy imparted to the particle as work done by the restoring force is *dissipated* or lost to the system responsible for exerting the damping force on it. For instance, for a particle executing a simple harmonic motion in a fluid, the damping force may arise due to viscous drag on it caused by the fluid (see sec. 7.5.8.3), in which case the energy is lost to the fluid, eventually heating it up.

Put differently, one has, from eq. (4-45),

$$(V_Q + K_Q) - (V_P + K_P) = E_Q - E_P = -\gamma \int_{t_1}^{t_2} v^2 dt, \quad (4-46)$$

which implies that the total energy of the particle (i.e., the sum of its potential and kinetic energies) *decreases* with time or, in other words, energy gets dissipated from the

oscillator, the rate of energy dissipation being given by

$$(\text{rate of energy dissipation}) \quad \mathcal{E} = -\frac{dE}{dt} = \gamma v^2. \quad (4-47)$$

If the damping constant b is small compared to the undamped frequency ω of the oscillator, then an approximate expression for the rate of energy dissipation is (refer to equations (4-42a), (4-42b))

$$\mathcal{E} = \gamma a^2 \omega^2 e^{-2bt} \sin^2(\omega t + \delta) \quad (b \ll \omega). \quad (4-48)$$

4.6 Forced SHM

In the sections above we have come across the concept of a restoring force proportional to the displacement of the particle, as also that of a damping or a retarding force proportional to the velocity. We shall now have a look at the effect of a *time-dependent* force on the motion of a particle undergoing a simple harmonic motion which, for the sake of generality, will be taken to be a damped one. In other words, we shall consider the motion of a particle under the action of *three* forces acting simultaneously on it - a restoring force, a retarding force, *and* a time-dependent force. The value of the time-dependent force at any given instant depends on the time t but not on the displacement or the velocity at that instant.

While the time-dependence can be of various different types, we shall consider, in particular, a *sinusoidal* time-dependence. In other words, we shall consider the effect of a force that undergoes a periodic change with some definite frequency and is given by an expression of the form

$$F = A \cos(pt). \quad (4-49)$$

According to what has been said above, the force itself is an instance of a physical quantity whose variation is a simple harmonic oscillation with angular frequency p (i.e., with frequency $\frac{p}{2\pi}$ and time period $\frac{2\pi}{p}$), and amplitude A .

One could equally well have taken an expression of the form $F = A \cos(pt + \delta)$ by including an initial phase δ , but the choice $\delta = 0$ does not imply any loss of generality.

The equation of motion of the particle is obtained by including this time-dependent force along with the restoring force and the retarding force in the right hand side of eq. (4-39) whereby one obtains

$$m \frac{d^2x}{dt^2} = -kx - \gamma \frac{dx}{dt} + A \cos(pt). \quad (4-50)$$

This equation provides a complete description of *forced simple harmonic motion* of a particle. At the same time, it describes the variation of physical quantities of other descriptions in numerous other situations like, for instance, the variation of current and voltage in an electrical circuit made up of an inductance, a capacitance, and a resistance, to which an *alternating EMF* is applied from an external source.

Analogous to the undamped or damped simple harmonic motion, one can determine the general solution of the above second order differential equation describing forced simple harmonic motion. Once again, the solution involves two constants which can be determined from the initial conditions of motion, whereby one arrives at a particular solution to the equation of motion. However, the *long term* behavior of the solution does *not* depend on these initial conditions. In other words, after a sufficiently large lapse of time from the instant which the initial conditions refer to, the motion does no longer depend on these initial conditions. Let me explain this in greater detail.

The general solution to eq. (4-50) can be expressed in the form a sum of *two* terms:

$$x(t) = \bar{x}(t) + x_0(t). \quad (4-51)$$

Here the expression for the first term $\bar{x}(t)$ is similar to the right hand side of eq. (4-42a) (we assume for the sake of concreteness that the motion is underdamped, i.e., $b < \omega$, where the parameters b and ω have been introduced above; the overdamped or critically damped cases can be described in similar terms) and it is this term that depends on the

initial conditions of motion. While it is relevant in describing the motion in the short run, its magnitude becomes progressively small for larger time intervals, eventually becoming negligible after a sufficiently large interval of time reckoned from the initial instant. A change in the initial conditions corresponds to a different expression for $\bar{x}(t)$ but the latter still becomes negligibly small after a sufficiently large lapse of time. One therefore refers to $\bar{x}(t)$ as the *transient* part of the solution to the equation of motion. In reality, the transient part of the solution is simply the same as the solution to eq. (4-39), i.e., the equation of motion *in the absence* of the periodic forcing.

The term $x_0(t)$ in eq. (4-51) is referred to as the *steady state solution* of the equation of motion (4-50) because, after a sufficiently large lapse of time when $\bar{x}(t)$ tends to zero, it is $x_0(t)$ that determines the variation of the displacement of the particle with time t . The expression for $x_0(t)$ turns out to be of the form (a convenient derivation of the solution is based on the use of *complex representation* of sinusoidally varying quantities, and will be found in section 13.5 in the context of an L-C-R circuit connected to an AC source of EMF)

$$x_0(t) = \frac{A}{m\sqrt{(\omega^2 - p^2)^2 + 4b^2p^2}} \cos(pt - \theta), \quad (4-52a)$$

where $\omega = \sqrt{\frac{k}{m}}$ and $b = \frac{\gamma}{2m}$ as before, and

$$\tan\theta = \frac{2bp}{\omega^2 - p^2}. \quad (4-52b)$$

Note that, in the absence of damping and external forcing, the particle executes a simple harmonic motion with angular frequency ω , where $\nu = \frac{\omega}{2\pi}$ is referred to as the *natural frequency* of the oscillation. The parameter b which has been referred to above as the damping constant, gives a measure of the degree of damping. Finally, we turn to the significance of the parameter θ .

As mentioned above, the expression for the displacement $x(t)$ of the particle effectively reduces to $x_0(t)$ after a sufficiently long time measured from the initial instant, when the transient part $\bar{x}(t)$ tends to zero, and the remaining part ($x_0(t)$) determines the steady

state motion of the particle. According to eq. (4-52a) above, this steady state motion is also a simple harmonic one in the sense of corresponding to a sinusoidal time variation, but one with a frequency $\frac{p}{2\pi}$ instead of the natural frequency $\frac{\omega}{2\pi}$. Moreover, the phase of this steady state simple harmonic motion at any given instant t is $(pt - \theta)$, which differs from the phase pt of the time dependent forcing term $A \cos(pt)$, the difference between the two phases being θ .

In other words, the variation of the displacement in the steady state motion occurs with the same frequency as the time-dependent force, but with a different phase. For instance, the displacement does not attain its maximum magnitude at the same instant as the force does. Again, the instant at which the force reverses its direction ($\cos(pt) = 0$), is not the same as the instant when the displacement gets reversed. The *phase lag* of the displacement with reference to the time dependent force is represented by θ . The phase lag attains the value $\frac{\pi}{2}$ when the forcing frequency p equals ω , the natural frequency.

The amplitude of the steady state simple harmonic motion is given by

$$B = \frac{A}{m\sqrt{(\omega^2 - p^2)^2 + 4b^2p^2}}. \quad (4-53)$$

This expression for the amplitude of the steady state motion depends on the angular frequency p of the time-dependent force. If the variation of B with p is considered for fixed values of the natural frequency ω and the damping constant b , then the nature of the variation will look as in the graph of fig. 4-14. The graph shows a maximum amplitude at the frequency $p = p_0$ given by

$$p_0 = \sqrt{\omega^2 - 2b^2}. \quad (4-54)$$

One observes that the frequency ($p = p_0$) at which the amplitude of steady motion attains a maximum value differs, in general, from the frequency ($p = \omega$) at which the phase lag θ attains the value $\frac{\pi}{2}$. Of these, the former is commonly referred to as the *resonant frequency* of the periodically forced simple harmonic motion, and the phenomenon of

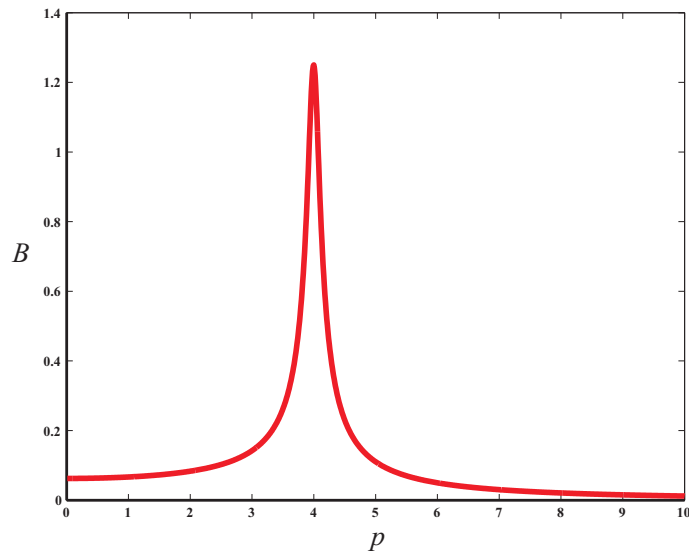


Figure 4-14: Variation of the amplitude (B) in the steady state motion ($x_0(t)$) in a forced SHM; the variation is shown as a function of the angular frequency p , for fixed b and ω ; parameters chosen: $A = 1$, $m = 1$, $\omega = 4.0$, $b = 0.1$ in arbitrary units; the amplitude is a maximum at the frequency $p = p_0$ given by the expression (4-54).

the amplitude of steady motion attaining a relatively large value at the value $p = p_0$ of the forcing frequency is referred to as *resonance* in the forced SHM.

If, however, the damping constant b satisfies $b \ll \omega$ then the two frequencies mentioned above become equal in an approximate sense and the term ‘resonant frequency’ then applies to either of the two.

4.6.1 Energy exchange in forced SHM. Resonance.

I now turn to a consideration of *energy exchange* in forced SHM because this helps us achieve a better understanding of the motion of the particle. Here one has to take into account the work done, on the particle under consideration, by *three* forces: the restoring force, the retarding force, *and* the time-dependent sinusoidal force.

1. More precisely, the system that exerts the time-dependent force on the particle performs work on it. At the same time, the particle performs work on the system that exerts the retarding force, which commonly happens to be the medium in which the particle moves or a body providing frictional resistance. A considerable

part of this work appears as the internal energy of the medium.

2. Recall that the system under consideration need not actually be a particle. For instance, it may even be an extended body of which the center of mass motion is described by an equation of the form (4-50).

Considering a complete period (say, τ , $\tau = \frac{2\pi}{p}$) of the time dependent force, one can calculate the average rate at which this force performs work on the particle and thereby supplies energy to it. On the other hand, the average rate at which the particle performs work against the retarding force, and thereby *loses* energy can also be worked out. For a simple harmonic motion with amplitude B given by eq. (4-53), the two turn out to be *equal* to each other, as a result of which the particle oscillates with a constant energy and hence a constant amplitude for an indefinite time. This explains why the *steady state* motion occurs with an amplitude B .

For an amplitude less than B , the particle gains in energy on the average and hence the amplitude, instead of remaining constant, increases with time. On the other hand, for an amplitude greater than B there occurs a net loss of energy on the average, leading to a decrease of the amplitude towards the value B . These two situations correspond to transient motions of the particle which eventually give way to the steady state motion with amplitude B .

The maximum value of the amplitude at, occurring at resonance ($p = p_0$), is given by (refer to eq. (4-53))

$$B_{\text{resonance}} = B(p_0) = \frac{A}{2mb\sqrt{(p_0^2 + b^2)}} = \frac{A}{\gamma\omega'}, \quad (4-55)$$

where the retardation constant γ and the damped frequency ω' have already been defined.

Evidently, the resonant amplitude increases with decreasing values of the damping constant b . For a small value of the damping constant, the energy supplied by the time-dependent force can be balanced by the damping loss only if the amplitude is large. In numerous situations of interest, the damping constant happens to be so small that the

resonant amplitude becomes very large. Such a large amplitude is commonly looked upon as the characteristic feature of resonance in forced oscillations.

Fig. 4-15 depicts graphically the variation of displacement against time in a forced SHM, where the variation of the time-dependent force is also shown in the figure for the sake of comparison. One notes from the graphs that the variation of displacement follows an apparently irregular pattern up to a certain time, after which the displacement varies sinusoidally with the same frequency as the force, but with a phase difference. It is this phase difference that has been indicated in eq. (4-52b). During the first phase the displacement involves a superposition of the transient part $\bar{x}(t)$ and the steady part $x_0(t)$ where the two are characterized by different frequencies, as a result of which the graph looks somewhat irregular. Afterwards, however, the transient part dies down and the steady sinusoidal variation appears.

Fig. 4-16 depicts the special case of *resonance*, where the forcing frequency ($\frac{p}{2\pi}$) is close to the natural frequency ($\frac{\omega}{2\pi}$), comparing the steady state motion with the case where the forcing frequency differs from the natural frequency. One observes here that the steady state amplitude is quite large in the case of resonance as compared to the case where resonance does not occur.

Instances of resonant oscillations and vibrations are commonly observed in a large number and variety of situations. A child pushing a swing carrying a friend of hers knows from experience that the swing attains a large amplitude when pushed at a particular rhythm. Similarly, when an alternating voltage of a particular frequency is applied to a circuit made up of an inductance, a capacitance, and a resistance, the current in the circuit oscillates with a large amplitude.

Problem 4-13

With reference to eq. (4-39) describing the damped SHM of a particle, consider the following values of the parameters (all in SI units): mass (m) = 0.001, force constant (k) = 10, retardation constant (γ) = 0.02. Calculate the proportional difference between the natural frequency and the damped

frequency. If the particle is acted upon by a periodically varying force with amplitude $A = 5$, calculate the resonant frequency (p_0), amplitude of steady motion at the resonant frequency, and the amplitude of steady motion at half the resonant frequency.

Answer to Problem 4-13

HINT: (All quantities are referred to in SI units) The natural frequency is $\omega = \sqrt{\frac{k}{m}} = 100$, while the damping constant is $b = \frac{\gamma}{2m} = 10$ (see formulae (4-40a), (4-40b)). The damped frequency is then (refer to formula (4-42b)) $\omega' = \sqrt{\omega^2 - b^2} = 99.5$ (approx.); the proportional difference between ω and ω' is $\frac{|\omega' - \omega|}{\omega} = .005$, i.e., 0.5%. The resonant frequency in the motion under the periodic force is $p_0 = \sqrt{\omega^2 - 2b^2} = 99$ (approx.; formula (4-54)). With $A = 5$ in formula (4-55), the amplitude of steady motion at resonance is seen to be $B = \frac{A}{\gamma\omega'} = 2.5$ (approx.). Finally, the amplitude at half the resonant frequency is seen from formula (4-53) to be 0.22 (approx.), showing that the amplitude varies rather sharply with forcing frequency.

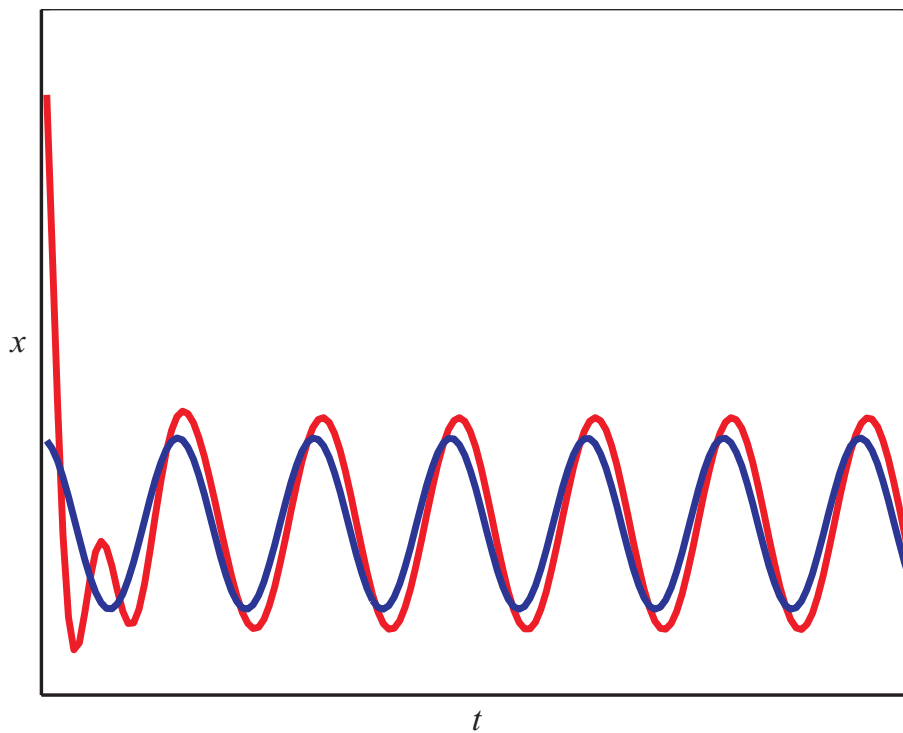


Figure 4-15: Variation of displacement (x) with time (t) in forced SHM; the displacement varies in a somewhat irregular manner up to a certain time after which the transient motion dies down and there occurs a steady state motion with sinusoidal time variation; the sinusoidal variation of the forcing term is also shown; the frequency of steady state motion is the same as that of the forcing term, while its phase differs from that of the latter; the parameters A , B , b , ω , p as also the initial conditions have been chosen arbitrarily, depicting only the nature of the variation; for these parameters, the phase difference θ is seen to be small, though discernible.

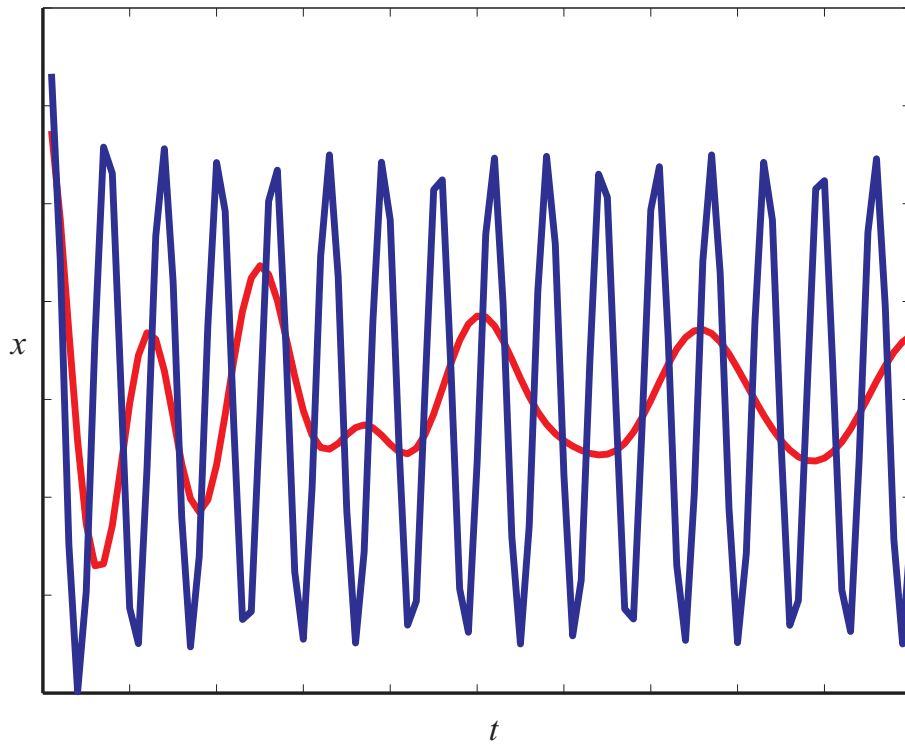


Figure 4-16: Variation of displacement (x) with time (t) in forced SHM for two different frequencies of the forcing term; when the forcing frequency differs to a considerable degree from the natural frequency of the oscillator, the amplitude of the steady state motion is comparatively small; if, on the other hand, the forcing frequency equals the natural frequency, resonance takes place and the amplitude of steady state motion becomes large.

Chapter 5

Gravitation

5.1 Introduction: Newton's law of gravitation

Gravitation is a fundamental and universal phenomenon in nature. It stands for an attractive force between all bodies in the universe, regardless of their size and mutual separation. It is gravitation that explains an immense variety of phenomena and motions we observe.

The basic principle in gravitation is expressed by *Newton's law of gravitation*. It starts by specifying the force of gravitation between two *particles* (idealized bodies having mass but no extension), and then supplements this with the *principle of superposition* that gives us the formula for the gravitational force between any two bodies, or even the force experienced by any given body due to any number of other bodies. The first part in this scheme of things, namely the force between any two particles, can be expressed as follows:

Every particle in the universe attracts every other particle by a force proportional to the product of the masses of the two particles and inversely proportional to the square of the distance between them, acting along the line joining them.

Suppose that there are two particles, which we label '1' and '2', with masses m_1 and m_2 respectively. If the separation between the particles be r then the magnitude of the

gravitational force between them is given by

$$F = G \frac{m_1 m_2}{r^2}. \quad (5-1)$$

Here G is a constant, referred to as the *universal constant of gravitation* (or, *gravitational constant* in brief). Its value in the SI system is

$$G = 6.67 \times 10^{-11} \text{ N} \cdot \text{m}^2 \cdot \text{kg}^{-2} \text{ (approx.)}. \quad (5-2)$$

As for the direction of the force, it acts along the line joining the two particles and is always an *attractive* one.

These facts can be combined into a single equation involving vectors. Let \mathbf{r}_1 and \mathbf{r}_2 be the position vectors of the two particles relative to any chosen origin and $\mathbf{r}_{12} = \mathbf{r}_1 - \mathbf{r}_2$ denote the position vector of the particle ‘1’ relative to particle ‘2’ (similarly, $\mathbf{r}_{21} = \mathbf{r}_2 - \mathbf{r}_1$ stands for the position vector of particle ‘2’ relative to particle ‘1’). Then the gravitational force \mathbf{F}_{12} on ‘1’ exerted by ‘2’ is given by the expression

$$\mathbf{F}_{12} = -G \frac{m_1 m_2}{r_{12}^3} \mathbf{r}_{12}. \quad (5-3a)$$

Here $r_{12} = |\mathbf{r}_1 - \mathbf{r}_2|$ is the magnitude of the distance between the two particles, which can also be written as r_{21} .

The force \mathbf{F}_{21} exerted by ‘1’ on ‘2’ can be expressed in a similar manner and satisfies (see fig. 5-1)

$$\mathbf{F}_{21} = -\mathbf{F}_{12}. \quad (5-3b)$$

These equations tell us that the forces exerted by the two particles on each other are *central* in nature, i.e., it act along the line joining the particles and are, moreover, equal and opposite, conforming to *Newton’s third law* (see sec. 3.17.2).

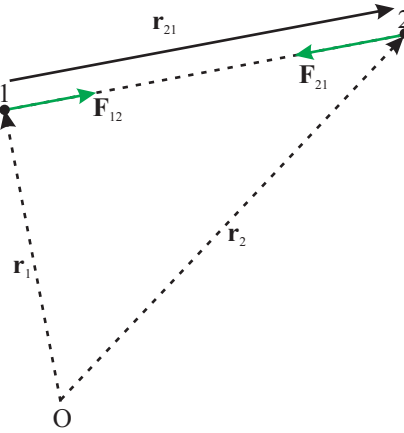


Figure 5-1: Illustrating Newton's law of gravitational force between two particles labeled 1 and 2; the force between the particles is central in nature, i.e., acts along the line joining the particles and is attractive; the position vectors of the two mass-points relative to a chosen origin O are respectively \mathbf{r}_1 and \mathbf{r}_2 , while \mathbf{r}_{21} is the position vector of '2' relative to '1'; \mathbf{F}_{12} and \mathbf{F}_{21} are forces on '1' and '2' respectively.

5.1.1 Principle of superposition

As I have mentioned above, Newton's law of gravitation is, in reality, a package comprising of two parts - the formula for the force between two mass points, and the principle of superposition. It is the latter to which I now turn.

Suppose that, instead of just two particles, we have *three* particles labeled as, say, 1, 2, and 3, interacting with one another by means of the gravitational force. What will then be the force experienced by any one of these, say '1', due to the other two, i.e., '2' and '3'? The principle of superposition states that this force, which we denote by \mathbf{F}_1 in the present instance is the sum of two terms,

$$\mathbf{F}_1 = \mathbf{F}_{12} + \mathbf{F}_{13}, \quad (5-4a)$$

where \mathbf{F}_{12} and \mathbf{F}_{13} (see fig. 5-2) are the forces exerted on '1' by '2' and '3' respectively, each regardless of the other, in accordance with the first part of Newton's law stated above. Thus, if \mathbf{r}_1 , \mathbf{r}_2 , and \mathbf{r}_3 be the position vectors of the three particles with respect

to any chosen origin and m_1, m_2, m_3 their respective masses, then

$$\mathbf{F}_1 = G \left[\frac{m_1 m_2}{|\mathbf{r}_2 - \mathbf{r}_1|^3} (\mathbf{r}_2 - \mathbf{r}_1) + \frac{m_1 m_3}{|\mathbf{r}_3 - \mathbf{r}_1|^3} (\mathbf{r}_3 - \mathbf{r}_1) \right]. \quad (5-4b)$$

More generally, for a system of N particles with masses m_1, m_2, \dots, m_N at locations given by position vectors $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N$ with respect to any chosen origin, the force of gravitational interaction on any one particle, say the i th one ($i = 1, 2, \dots, N$) due to all the others is given by the expression

$$\mathbf{F}_i = \sum_{j(j \neq i)} G \left[\frac{m_i m_j}{|\mathbf{r}_j - \mathbf{r}_i|^3} (\mathbf{r}_j - \mathbf{r}_i) \right], \quad (5-4c)$$

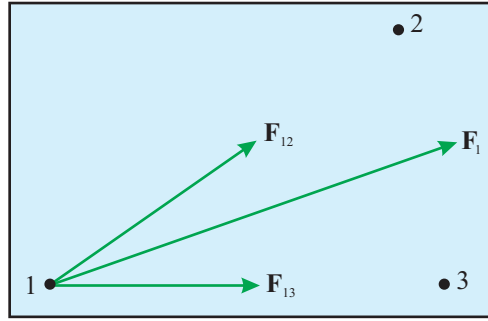


Figure 5-2: Illustrating the principle of superposition in gravitation; the force \mathbf{F} on the particle 1 exerted simultaneously by particle 2 and particle 3 is the vector sum of forces \mathbf{F}_{12} and \mathbf{F}_{13} exerted by '2' and '3' respectively, one in the absence of the other.

where the summation is to be carried out for all j from 1 to N , excluding the value $j = i$.

One can determine the force exerted by one extended body on another in a similar manner. Thus, considering any one of the two bodies as a collection of particles, one can work out the force on any one of those particles exerted by all the particles making up the other body as in eq. (5-4c), and then take the vector sum of the forces on all the particles in the first body. The force on the second body can also be worked out similarly.

Newton's law of gravitation along with the principle of superposition implies that the

gravitational force on a particle due to one or more fixed bodies is proportional to the mass of the particle, which is a fact of considerable significance. In particular, it underlines a fundamental similarity between the gravitational force and inertial forces that arise due to the acceleration of a frame of reference. As we saw in sections 3.10.3, 3.21, the inertial forces acting on a particle are all proportional to its mass, as a result of which its acceleration is independent of the mass. This similarity between the gravitational force and inertial forces is indicative of a special feature of the gravitational force as compared with other forces observed in nature and implies that gravitation may have a significance not shared by these other forces. Indeed, the idea underlying the definition of inertial frames finds an extension in the *general theory of relativity* (see chapter 17 for a brief introduction) where gravitation is to be taken into account in this definition.

Problem 5-1

Find the vector expression for the force of gravitation on a particle of mass (all SI units implied) $m = 0.015$ located at $\mathbf{r} = 2\hat{i} - \hat{j} + \hat{k}$, due to two other particles, one of mass $m_1 = 0.03$ at $\mathbf{r}_1 = 2\hat{i} + 2\hat{j} + \hat{k}$ and the other of mass $m_2 = 0.05$ at $\mathbf{r}_2 = -\hat{i} - \hat{j} - 3\hat{k}$.

Answer to Problem 5-1

HINT: The vector separation of the first particle from the second and third particles are respectively $\mathbf{u}_1 = -3\hat{j}$ and $\mathbf{u}_2 = 3\hat{i} + 4\hat{k}$. The force is given by $\mathbf{F} = -Gm\left(\frac{m_1}{u_1^3}\mathbf{u}_1 + \frac{m_2}{u_2^3}\mathbf{u}_2\right) = -G\left(\frac{45}{27}(-3\hat{j}) + \frac{75}{125}(3\hat{i} + 4\hat{k})\right) \times 10^{-5} = G\left(-\frac{9}{5}\hat{i} + 5\hat{j} - \frac{12}{5}\hat{k}\right) \times 10^{-5}$, where all quantities are in SI units.

Starting from Newton's law of gravitation as expressed by the above formulas, I will now introduce the concepts of gravitational field, gravitational intensity, and gravitational potential.

5.2 Gravitational intensity and potential

5.2.1 Gravitational intensity

Fig. 5-3 depicts a point mass m located at P, due to which a force is exerted on another point mass m' located at R. Since this second mass m' and the point R can be chosen arbitrarily, one can say that the mass m located at P creates a certain *influence* around itself by virtue of which it exerts a force on a second mass located at any point such as R (the force on the mass m exerted by the second mass m' is not relevant in the present context).

If a *standard* or *reference* mass is placed at any such point, say, at R, then the force on that mass can be taken as a quantitative measure of the influence at R set up by the mass m at P. If this reference mass m' at the point R is taken to be unity, then the force on it exerted by the mass m is referred to as the *gravitational intensity* at R due to the mass m at P.

In other words, the gravitational intensity is the force on a unit mass placed at the point under consideration. Evidently, the gravitational intensity is a vector quantity, which can be obtained in the present instance from eq. (5-3a) on putting $m_1 = 1$ (kg, in the SI system), $m_2 = m$, and, say, $\mathbf{r}_{12} = \mathbf{r}$, the position vector of the point R relative to P. Denoting this by \mathbf{I} one obtains

$$\mathbf{I} = -\frac{Gm}{r^3}\mathbf{r}. \quad (5-5a)$$

Alternatively, assuming that the mass m is located at a point P with position vector \mathbf{r} relative to a chosen origin, the gravitational intensity at a second point R with position vector \mathbf{r}' is given by the expression

$$\mathbf{I} = -Gm \frac{\mathbf{r}' - \mathbf{r}}{|\mathbf{r}' - \mathbf{r}|^3}. \quad (5-5b)$$

It helps to write this in the simpler form

$$\mathbf{I} = -\frac{Gm}{u^2}\hat{u}, \quad (5-5c)$$

where u stands for the distance from the *source point* P to the *field point* R,

$$u = |\mathbf{r}' - \mathbf{r}|, \quad (5-5d)$$

and \hat{u} is the unit vector directed from the former to the latter

$$\hat{u} = \frac{\mathbf{r}' - \mathbf{r}}{|\mathbf{r}' - \mathbf{r}|}. \quad (5-5e)$$

The gravitational intensities at various different points like R due to the source mass m (i.e., the mass whose gravitational influence is under consideration) located at P, make up a *field* of intensities or, equivalently, a *gravitational field*. It is the gravitational field that describes completely the influence set up by the mass m at various points in space.

Knowing the intensity \mathbf{I} at a field point R, the force on a mass, say m' placed at the field point can be obtained as (recall the definition of intensity as the force on a unit mass) $\mathbf{F} = m'\mathbf{I}$.

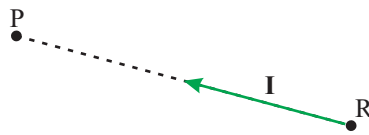


Figure 5-3: Illustrating the concept of gravitational intensity; a point mass m at the source point P generates a field around it, the intensity at the field point R being \mathbf{I} ; the intensity represents the force on a particle of unit mass imagined to be located at R.

Now imagine a number of source masses m_1, m_2, \dots, m_N located at points with position vectors, say, $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N$ respectively, instead of a single source mass m at \mathbf{r} . Together, all these masses set up a gravitational field in space, which is described in terms of the gravitational intensities at the various field points. Considering a typical field point, say \mathbf{r} , the intensity at this point is defined as the force on a unit mass imagined to be placed

at \mathbf{r} . Making use of Newton's law expressed by eq. (5-4c), the intensity is found to be

$$\mathbf{I} = -G \sum_{i=1}^N \frac{m_i(\mathbf{r} - \mathbf{r}_i)}{|\mathbf{r} - \mathbf{r}_i|^3}. \quad (5-6a)$$

Correspondingly, the force on a mass m placed at the field point (this implies a slight change in notation, since the mass at the field point was previously denoted by m') is given by

$$\mathbf{F} = m\mathbf{I}, \quad (5-6b)$$

(check the above statements out).

Fig. 5-4 depicts schematically a gravitational field in a region R of space, where a vector \mathbf{I} , represented by a directed line segment, is associated with every point \mathbf{r} in the region. One thereby has a vector function of the vector variable \mathbf{r} which can be denoted by the symbol $\mathbf{I}(\mathbf{r})$. The vector function $\mathbf{I}(\mathbf{r})$ is given by an expression of the form (5-6a) for a field set up by a number of point masses. This is a particular instance of a *vector field* introduced in section 2.13, another instance of which is provided by an electrostatic field to be discussed in chapter 11.

Both a gravitational field and an electric field are instances of a *field of force* (see section 3.11). Indeed, considering any given mass, say, m , a field of force $\mathbf{F}(\mathbf{r})$ can be obtained from the gravitational field $\mathbf{I}(\mathbf{r})$ by using eq. (5-6b), this being the force experienced by the mass m at various points in the gravitational field, and the gravitational field is simply the field of force for $m = 1$.

Problem 5-2

Four identical massive particles, each of mass m , are located at the corners of a square of edge length a and revolve, under their mutual gravitational interaction, around their common center of mass on a fixed circular orbit such that the lines joining the instantaneous positions of the particles at any point of time form a similar square of edge length a , where the corners of the

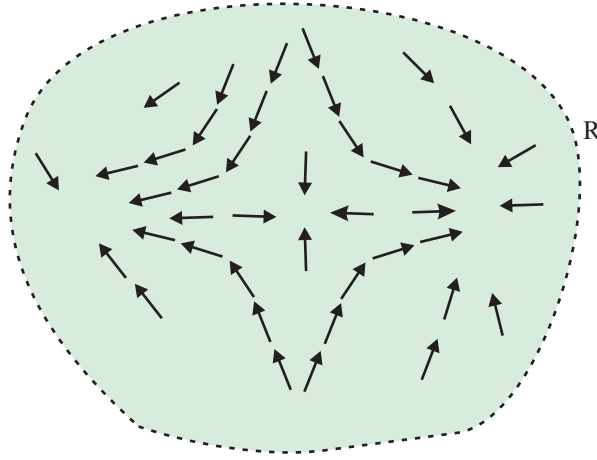


Figure 5-4: A gravitational field (schematic) set up in a region R (source masses not shown in the figure), represented by an intensity vector at every point in the region.

square lie on the circle. What should be the velocity of each particle for such a motion to be possible?

Answer to Problem 5-2

HINT: The resultant force on each particle due to the gravitational attraction of the other three particles is $\frac{1+2\sqrt{2}}{2} \frac{Gm^2}{a^2}$ directed towards the fixed center (check this out). This must provide for the centripetal force on the particle necessary for it to move along a circle of radius $\frac{a}{\sqrt{2}}$. Hence the velocity v of each of the particles satisfies the relation $\frac{mv^2}{\frac{a}{\sqrt{2}}} = \frac{1+2\sqrt{2}}{2} \frac{Gm^2}{a^2}$, from which one obtains $v = \frac{\sqrt{(4+\sqrt{2})}}{2} \sqrt{\frac{Gm}{a}}$.

5.2.2 Gravitational potential

An important feature of the force field $\mathbf{F}(\mathbf{r}) = m\mathbf{I}(\mathbf{r})$ mentioned above is that it is a *conservative* one. Correspondingly $\mathbf{I}(\mathbf{r})$ also constitutes a conservative vector field.

Recall from chapter 3 that a force field acting on a particle is said to be conservative if, given any two points P and Q in the field, the work done by the force in a displacement from Q to P along any given path turns out to be independent of the path, being determined solely by the positions of the two points P and Q . An equivalent way of stating this is to say that the line integral $\int_Q^P \mathbf{F} \cdot d\mathbf{r}$ evaluated along a path connecting

P and Q is independent of the path chosen. Recall further that, for a conservative force field, one can define a potential energy (V) of the particle where the difference of potential energies at P and Q (i.e., $V_P - V_Q$) is the work done *against* the force in a displacement from Q to P.

This means that the work done by the gravitational field in a displacement of a particle of mass m from any point Q to another point P along an arbitrarily chosen path is independent of the path followed. Further, a potential energy of the particle at any chosen point, say, P with position vector \mathbf{r} , can be defined as the work done against the field in a displacement of the particle from a chosen reference point to the point P under consideration. If now the mass m of the particle is chosen to be unity (1 kg in the SI system) then the potential energy of the mass is referred to as the gravitational *potential* (or, in brief, the potential) at the point \mathbf{r} . With $m = 1$, the force $\mathbf{F}(\mathbf{r})$ reduces to the gravitational intensity $\mathbf{I}(\mathbf{r})$ and thus the potential (commonly denoted by $V(\mathbf{r})$) is given by the formula

$$V(\mathbf{r}) = - \int_{\mathbf{r}_0}^{\mathbf{r}} \mathbf{I} \cdot d\mathbf{r}, \quad (5-7)$$

where \mathbf{r}_0 is the position vector of the chosen reference point and the integration is performed along any arbitrarily chosen path from the reference point to the point under consideration.

The potential so defined is undetermined to the extent of an additive constant since a different choice of the reference point causes a change in the value of $V(\mathbf{r})$ by a constant amount (check this out.)

The gravitational potential V being the potential energy of a unit mass at any given point in a gravitational field, the potential energy of a mass m is given by the expression

$$\mathcal{V} = mV, \quad (5-8a)$$

and the work done *against* the field in displacing the mass m from a point Q to another

point P is

$$W_{Q \rightarrow P} = m(V_P - V_Q). \quad (5-8b)$$

Since the unit of energy in the SI system is the joule (J), the unit of gravitational potential will be $\text{J} \cdot \text{kg}^{-1}$.

5.2.3 Gravitational potential: summary

I now summarize what I have said above regarding the potential in a gravitational field.

1. The field of force acting on a point mass in a gravitational field created by a given distribution of fixed masses is conservative in nature.
2. One can define a potential at any point in such a field with respect to a chosen reference point. If the position vector of the point be \mathbf{r} , then the potential $V(\mathbf{r})$ is the work done against the force due to the field in transferring a unit mass from the reference point to the point \mathbf{r} .
3. The potential energy of a mass m placed at the point \mathbf{r} is $mV(\mathbf{r})$.
4. If P and Q be any two points in a gravitational field, the work done against the gravitational force in transferring a mass m from Q to P along any chosen path is given by the expression (5-8b).
5. If a different reference point is chosen in defining the potential then the latter gets changed by the addition of a constant term, but this non-uniqueness of the potential does not show up in the potential *difference* between any two points since the constant additive term gets canceled in the potential difference.
6. For a given choice of the reference point, the potential *at* that point will evidently be zero.

5.2.4 'Potential at infinity'.

In determining the potential at any point in a gravitational field, the reference point is commonly taken to be 'at infinity'. What this specifically means is the following.

With any point O chosen as the origin, imagine a straight line to be drawn from O in any direction, and consider a point Q on the straight line. Now imagine the point Q to be shifted to ever greater distances from O along the straight line, eventually moving out to an infinitely large distance. Assume that the potential in the gravitational field is determined with Q chosen as the reference point, in the limit of Q being moved out to an infinite distance. However, such an approach would be meaningful only if the value of the potential so determined were independent of the direction along which the line OQ was chosen to start with, in which case Q would be termed *the point at infinity*, irrespective of the direction.

This requirement is ensured if all the source masses generating the gravitational field under consideration are located *in a finite region of space*. In that case, any point located at an infinitely large distance from the origin can serve as the ‘point at infinity’. If, however, some of the source masses themselves are spread out to infinite distances, one has to specify the *direction* of the line OQ along which the point at infinity is chosen. For any *other* direction then, the potential at an infinitely large distance would not be zero. Alternatively, a point located at some *finite* distance may be chosen as the reference point, in which case again, the ‘potential at infinity’ would, in general, be non-zero.

You will find this explained in some more detail in section 11.4.4 in connection with the *electrostatic* field.

5.2.5 Potential due to a point mass

As explained above, if all source masses are located within a finite distance from the origin, the reference point can be chosen to be at infinity, i.e., the potential at infinity can be assumed to be zero. For instance, consider first the simplest of situations, where there is only *one* source mass located at a point, say, P with position vector \mathbf{r} relative to a chosen origin O (see fig. 5-5). What will be the potential due to this mass at a field point, say, R , with position vector \mathbf{r}' ? By definition, the potential at R is the work done against the force exerted by the source mass (say, m) when a unit mass is brought to the position R from an infinite distance along any chosen line. In the present instance it is

convenient to choose this as the line joining P to R, imagined to be extended to infinity. One can then work out the expression for the work done against the force exerted on the unit mass by the source mass m , which turns out to be

$$V = -\frac{Gm}{|\mathbf{r}' - \mathbf{r}|}, \quad (5-9a)$$

or, more simply,

$$V = -\frac{Gm}{u}. \quad (5-9b)$$

In this last expression, u stands for the distance from the source point to the field point, as in eq. (5-5d).

Problem 5-3

Show that the gravitational potential due to a point mass is given by expression (5-9b), with notation introduced above.

Answer to Problem 5-3

Denoting by u' the distance from P to R' (see fig. 5-5), the work done against the force exerted by the source charge in displacing a unit charge from a distance u' to $u' + \delta u'$ is $-\mathbf{F}(u') \cdot \hat{u} \delta u'$, where \hat{u} is the unit vector pointing from P to R in the figure and where, according to eq. (5-5c), $\mathbf{F}(u') = -\frac{Gm}{u'^2} \hat{u}$. Breaking up the path (along which the unit charge is transferred) into a large number of small segments and summing up all these segments one gets the required potential. In the limit of the lengths of the segments going to zero, the required expression reduces to the integral

$$V = Gm \int_{\infty}^u \frac{du'}{u'^2}, \quad (5-10)$$

which gives formula (5-9b).

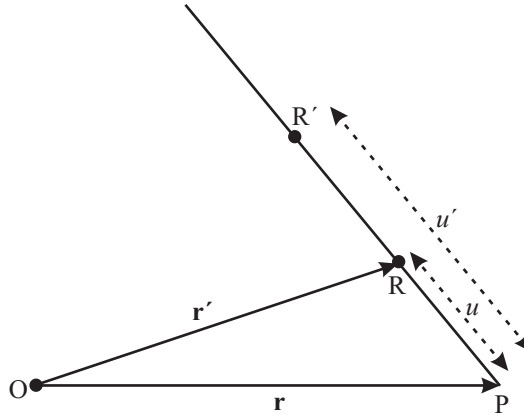


Figure 5-5: Potential due to a point mass; a source mass m at P produces a potential V at the field point R ; imagining the line PR to be extended out to infinity, the potential is the work done against the force exerted by the source mass in bringing a unit mass from infinite distance down to R along this line; R' is an intermediate point arrived at by the unit mass in this process; O is any chosen origin.

5.2.6 Potential due to a number of point masses

We will now obtain an expression for the potential at any point in the field set up by a number of source masses. Let N number of point masses m_1, m_2, \dots, m_N be located at points with position vectors $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N$ with respect to any chosen origin. According to the principle of superposition, the potential V at a field point with position vector \mathbf{r} will be the sum of terms V_i ($i = 1, 2, \dots, N$), where V_i is the potential at the field point due to the mass m_i located at \mathbf{r}_i independently of the other source masses (reason this out). Using for each source mass an expression of the form (5-9a), one gets

$$V_i = -G \frac{m_i}{|\mathbf{r} - \mathbf{r}_i|}, \quad (5-11a)$$

and

$$V = \sum_{i=1}^N V_i = -G \sum_{i=1}^N \frac{m_i}{|\mathbf{r} - \mathbf{r}_i|}. \quad (5-11b)$$

The following expression looks simpler:

$$V = -G \sum_{i=1}^N \frac{m_i}{u_i}. \quad (5-11c)$$

In this last expression u_i stands for the distance from the i th source point to the field point under consideration.

Problem 5-4

Two identical stellar bodies of spherical shape, each of mass m and radius a , are at rest in a certain inertial frame, being separated by a distance d . As they approach each other under their mutual gravitational attraction, what will be the velocity of either body, as measured in that frame, (a) when their separation is $\frac{d}{2}$, and (b) when they are about to collide?

Answer to Problem 5-4

HINT: The energy of the system made up of the two bodies is initially just their mutual potential energy since their kinetic energy is zero. The potential energy is $-G\frac{m^2}{d}$, being the potential energy (mass times gravitational potential) of either of the two bodies in the gravitational field of the other (the gravitational interaction of two homogeneous spherical bodies is equivalent to that two point masses; refer to sec. 5.3.3.3). When separated by a distance $\frac{d}{2}$, the potential energy gets reduced to $-G\frac{m^2}{\frac{d}{2}}$. Thus, from the principle of conservation of energy, the kinetic energy at separation $\frac{d}{2}$ will be $G\frac{m^2}{d}$. The velocity (v) of either body is thus to be obtained from the relation $2 \times \frac{1}{2}mv^2 = G\frac{m^2}{d}$. The separation becomes $2a$ when the bodies are about to collide. The decrease in potential energy is now $Gm^2\left(\frac{1}{2a} - \frac{1}{d}\right)$, which will then be equal to $2 \times \frac{1}{2}mv'^2$, where v' is the velocity of either body at the time of collision.

Problem 5-5

Three particles A, B, C, of masses m_1 , m_2 , and m_3 respectively are placed on the x-axis of a co-ordinate system at points $x = 0$, $x = d(> 0)$, and $x = l(> 2d)$. B is now shifted to $x = l - d$ with A and C held fixed. What is the work done by the force exerted by A on B, and by the net force of A and C on B? What is the increase in potential energy of B in the displacement?

Answer to Problem 5-5

HINT: The force exerted by A on B, when the latter is at a point with co-ordinate x ($d < x < l - d$), is $F_1 = -\frac{Gm_1m_2}{x^2}$, and the net force on B in this position is $F = -\frac{Gm_1m_2}{x^2} + \frac{Gm_2m_3}{(l-x)^2}$, where these

forces are along the x-axis, carrying their own signs. The work done by F_1 in a displacement of B from $x = d$ to $x = l - d$ is $\int_d^{l-d} F_1 dx = -Gm_1m_2 \frac{l-2d}{l(l-d)}$. The work done by the net force F is $\int_d^{l-d} F dx = Gm_2(m_3 - m_1) \frac{l-2d}{l(l-d)}$. The increase in the potential energy of B in the displacement, which is the work done *against* the net force, is $-Gm_2(m_3 - m_1) \frac{l-2d}{l(l-d)}$.

5.2.7 Describing a gravitational field

Note that the equations (5-6a) and (5-11b) are obtained from (5-5b) and (5-11a) respectively by invoking the superposition principle. However, while eq. (5-6a) involves a *vector* sum of contributions from individual source masses, eq. (5-11b) gives the potential as a sum of *scalar* contributions. As a result, the calculation of the potential at a point in a gravitational field is sometimes more convenient than that of the intensity, though, it is the intensity that is of more direct physical relevance.

Given the intensity field, one obtains the potential at a point r by evaluating the line integral in formula (5-7) along any appropriately chosen path, since the integral is independent of the path and depends only on r and on the reference point r_0 where a common choice for r_0 is 'the point at infinity'. The problem of deriving the intensity from the potential is the inverse one and will be addressed in chapter 11 in connection with the electric field. One way to state the result is the following: given any point in a gravitational field, the rate of change of the potential with distance along any chosen direction at that point, taken with the opposite sign, gives the component of the gravitational intensity in that direction. Equivalently, one can say that the gravitational intensity is the gradient of the potential, taken with a negative sign:

$$\mathbf{I} = -\nabla V, \quad (5-12)$$

where the symbol ∇ stands for the gradient (see section 2.14.1) and is explained at some length in section 11.4.7.

Having got at the concepts of gravitational intensity and potential, one can describe a gravitational field *geometrically* in terms of *lines of force*, an alternative geometrical description being possible in terms of *equipotential surfaces*. While the former approach

holds for any force field, the latter is a convenient description for a conservative one. Once again, you will find these ideas described in greater details in chapter 11.

5.3 Gauss' principle in gravitation

5.3.1 Flux of gravitational intensity

In fig. 5-6 below, S is a closed surface in a gravitational field, on which P is any chosen point. A small area around P lying on S has been shown in the figure, which can be assumed to be a plane one, lying in the tangent plane at P to the surface. Also shown are the outward drawn normal (PN) at P (i.e., the normal to the surface, pointing away from its interior) and the gravitational intensity (\mathbf{I}) at P , represented by the directed line segment PR.

If the area of the small element around P be δs and the unit vector along the normal PN be \hat{n} , then the vector area (see section 2.6.2) of the element will be $\vec{\delta s} = \delta s \hat{n}$. The expression $\mathbf{I} \cdot \vec{\delta s}$ is then referred to as the *flux* of the gravitational intensity through the element of area under consideration. Evidently, this will be a small quantity in the present context. Denoting this by $\delta\Phi$ one has

$$\delta\Phi = \mathbf{I} \cdot \vec{\delta s} = I \delta s \cos \theta, \quad (5-13)$$

where θ is the angle between \mathbf{I} and \hat{n} .

Imagining the entire closed surface S to be divided into a large number of such small area elements, and obtaining the flux through each such element in the above manner one may finally work out the sum of all these small quantities so as to arrive at the *total* flux through the closed surface S :

$$\Phi = \sum \delta\Phi = \sum \mathbf{I} \cdot \vec{\delta s}, \quad (5-14)$$

where the summation is over all the small area elements on S . If, now, the area of each

of the elements be assumed to be vanishingly small (i.e., to go to zero), then the flux is seen to reduce to the *surface integral* (see section 2.14.2) of the gravitational intensity on S:

$$\Phi = \oint \mathbf{I} \cdot d\vec{s} = \oint I \cos \theta ds. \quad (5-15)$$

Here the symbol \oint indicates the surface integral on a closed surface.

If the surface under consideration is not a closed one, than one has first to choose one of the two sides of the surface for defining the direction of the unit vector \hat{n} , and then evaluate a sum of the form (5-14), where the area elements now cover the surface under consideration and where, once again, the area of each element is to be made vanishingly small. In this case the expression $\oint \mathbf{I} \cdot d\vec{s}$ in formula (5-15) will have to be replaced with $\int \mathbf{I} \cdot d\vec{s}$.

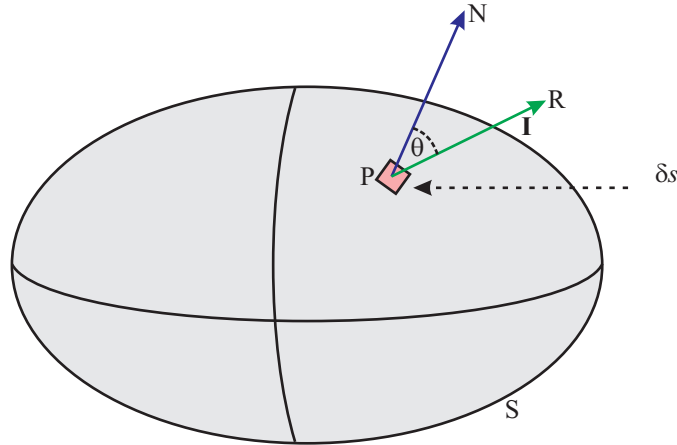


Figure 5-6: Illustrating the concept of gravitational flux over a surface, the surface chosen being a closed one in the present instance; a small area element on the surface is shown at the point P, where the outward drawn normal PN and the gravitational intensity I are also shown; the flux through the area element is $\mathbf{I} \cdot d\vec{s}$; the flux over the entire closed surface is obtained by summing up such contributions.

5.3.2 Gauss' principle

Evidently, the value of flux over any closed surface in a gravitational field will depend on the source masses responsible for the setting up of the field, where the closed surface may be an imagined one rather than the boundary surface of a material body. However, the relation between the source masses and the flux over a closed surface is a curious one, and forms the content of what is referred to as Gauss' principle in gravitation.

Recall from eq. (5-6a) that the gravitational intensity at any chosen point in a gravitational field is determined by the values and locations of *all* the source masses setting up the field. The *flux over a closed surface*, however, is determined *only* by the *total mass within* that closed surface. Denoting the latter by the symbol M , the flux is given by

$$\oint \mathbf{I} \cdot d\vec{s} = -4\pi GM. \quad (5-16)$$

This relation between the gravitational flux over a closed surface and the total mass located within that surface is the mathematical expression of Gauss' principle.

Notice that the gravitational flux over a closed surface does not depend on the values and locations of the source masses external to that surface, nor does it depend on the *locations* of the source masses in the interior, depending instead on the total value of the interior masses alone. This, then, constitutes the content of Gauss' principle.

Gauss' principle is often a useful and convenient means for determining the intensities in gravitational fields created by *symmetric distributions* of source masses. One can also determine the intensity in a gravitational field by applying Newton's law of gravitation as expressed by eq. (5-6a). However, Gauss' principle is at times a more convenient one in this respect.

Once again, much of what has been said here carries over almost verbatim to what I will have to say in chapter 11 in the context of an electrostatic field. In particular, the derivation of Gauss' principle in gravitation is entirely similar to that of Gauss' principle in electrostatics (see sec. 11.8.2.3).

5.3.3 Application: a spherically symmetric body

Fig. 5-7 ((A), (B)) shows a body B with a *spherically symmetric* mass distribution which means that, with the origin O chosen at the center of mass of B, the density at any point, say, R with position vector \mathbf{r}' depends only on $|\mathbf{r}'|$, i.e., on the distance OR, and *not* on the *orientation* of the line OR. Consider now a field point P in the gravitational field of this body, where P may be located either in the exterior (fig. 5-7(A)) or in the interior (fig. 5-7(B)) of the body. In determining the gravitational intensity at P, it is convenient to apply Gauss' principle.

Consider, for this purpose, an imagined spherical surface S, with center at O, and passing through P. Such a surface, on which the flux of gravitational intensity will be evaluated and equated to $-4\pi G$ times the mass contained in its interior, is referred to as a *Gaussian surface* in the given context.

For the spherical Gaussian surface specified above, the gravitational intensity has to be everywhere directed radially (either inward or outward), the magnitude of the intensity being, moreover, the same at all points - a consequence of the spherical symmetry of the mass distribution of the body under consideration.

Let us denote the *outward* radial intensity at any point on the Gaussian surface by I where I is to carry its own sign (if, at the end of the exercise, I turns out to be negative, then that will mean that the intensity is, in reality, directed *inward* at all points; reason out why I should indeed turn out to be negative or, at most, zero). Since the outward drawn normal at any point like P on S is also directed radially one has, in eq. (5-15), $\theta = 0$. Since, moreover, I is the same throughout the surface S, the flux works out to IA , where A , the area of the Gaussian surface, is $4\pi r^2$ (r stands for the distance OP). Gauss' principle, eq. (5-16), then gives

$$4\pi I r^2 = -4\pi G M(r), \text{ i.e., } I = -G \frac{M(r)}{r^2}, \quad (5-17a)$$

where $M(r)$ is the mass of the body B contained within the Gaussian surface S. For

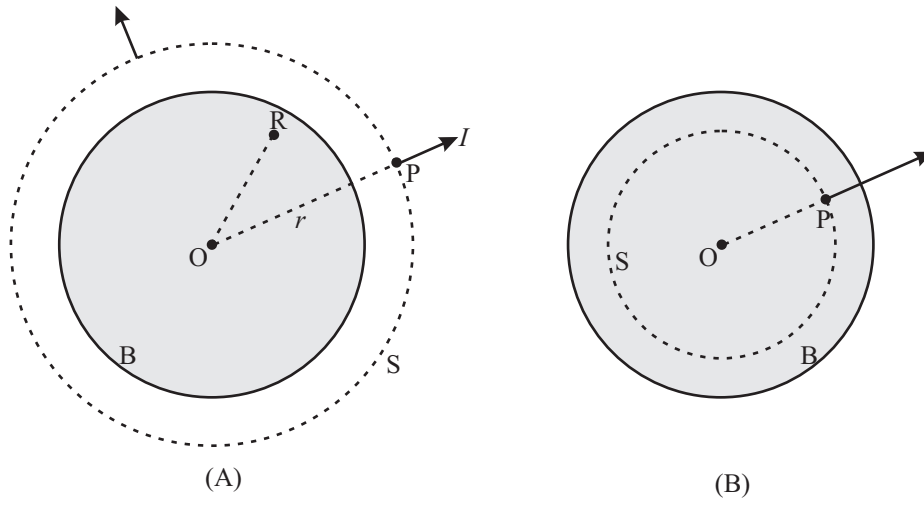


Figure 5-7: Illustrating the application of Gauss' principle to the calculation of intensity at a point P due to a body B with spherically symmetric mass distribution centred at O; the density at any point R in the body depends only on the distance OR and not on the direction in which the vector $\mathbf{r}' (= \vec{OR})$ points; P is a field point where, in (A), it lies outside the body B and, in (B), in the interior of B; S is an imagined spherical surface through P (the *Gaussian surface*); the intensity \mathbf{I} at P is directed radially and is of the form $\mathbf{I} = I \frac{\mathbf{r}}{r}$, where I is constant throughout S; the intensity vector at one other point on S is shown in (A); in the end, I turns out to be negative and so the intensity is, in reality, directed inward.

the situation shown in fig. 5-7(A), $M(r)$ is the mass of B itself, while in fig. 5-7(B), $M(r)$ represents that part of the mass of B that falls inside the Gaussian surface of radius r .

Equivalently, one can express, in vectorial form, the gravitational intensity at the point P as

$$\mathbf{I} = -\frac{GM(r)}{r^3} \mathbf{r}, \quad (5-17b)$$

where \mathbf{r} denotes the position vector of P relative to O.

It is worthwhile to look at the following particular cases.

5.3.3.1 Intensity and potential at an external point

For an external point (fig. 5-7(A)), $M(r)$ is nothing but the total mass (M) of the body B, being thus independent of r , the distance of the field point from the center. One

therefore has

$$I = -\frac{GM}{r^2}, \quad \mathbf{I} = -\frac{GM}{r^3} \mathbf{r}. \quad (5-18a)$$

Comparing with eq. (5-5a), the intensity due to a body with spherically symmetric mass distribution at an external point is seen to be the same as the intensity that would result if the entire mass of the body were concentrated at the center O . This immediately gives the following result for the potential at the external point P :

$$V = -\frac{GM}{r}, \quad (5-18b)$$

the gravitational potential *energy* of a particle of mass, say, m placed at P being then

$$\mathcal{V} = -\frac{GMm}{r}. \quad (5-18c)$$

5.3.3.2 A spherical shell

Fig. 5-8 shows a *homogeneous* spherical shell of mass M , and inner and outer radii a and b respectively, and a field point P at a distance r from the center O of the shell where, in (A), P lies outside the shell ($r > b$), in (B) P lies inside the material of the shell ($a < r < b$), and in (C), P lies in the hollow interior of the shell ($r < a$).

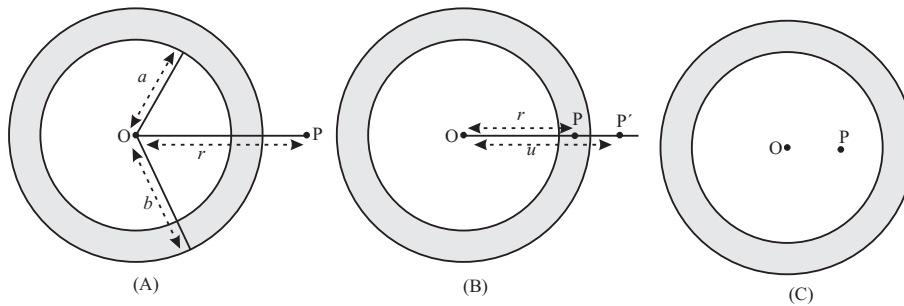


Figure 5-8: Calculating the intensity and potential due to a spherical shell of inner and outer radii a and b ; (A) field point P external to the shell, (B) field point inside the material of the shell, (C) field point in the hollow interior of the shell; in (B) the integration for determining the potential is performed along the radial line from infinity down to the point P , where P' is an intermediate point at a distance u from O ; in (C), the intensity is zero, while the potential has a constant value everywhere in the hollow, equal to that at $r = a$.

In order to determine the gravitational intensity in any of these three cases, a Gaussian surface S is to be imagined in the form of a spherical surface with O as center, passing through the field point P . In the case (A), the intensity and potential at P are already given by (5-18a) and (5-18b). In the case (B), only a part of the mass of the spherical shell lies inside the Gaussian surface, giving

$$M(r) = \frac{M(r^3 - a^3)}{b^3 - a^3}, \quad (5-19a)$$

and hence

$$I = -GM \frac{r^3 - a^3}{r^2(b^3 - a^3)}, \quad \mathbf{I} = -GM \frac{r^3 - a^3}{r^3(b^3 - a^3)} \mathbf{r} \quad (a < r < b), \quad (5-19b)$$

where \mathbf{r} stands for the radius vector from O to P (check these results out).

It is now an interesting exercise to work out the *potential* at P with $a < r < b$. Considering the straight line OP extended to infinity, and any point P' with a position vector, say, \mathbf{u} on this line, one has to invoke here eq. (5-7) where the integration is to be performed (with \mathbf{u} replacing the integration variable \mathbf{r} so as not to mix up symbols) along this straight line, starting from an infinite distance up to the point P . Writing $\mathbf{I} = I(u)\hat{\mathbf{u}}$ and $d\mathbf{u} = du\hat{\mathbf{u}}$, where $\hat{\mathbf{u}}$ is the unit vector along OP' , one obtains, for the potential at P ,

$$V(r) = - \int_{\infty}^r I(u) du. \quad (5-20)$$

The important thing to note is that the integration here *cannot be performed at one go*, because the expression for $I(u)$ in the range $b < u < \infty$ (refer to eq. (5-18a)) differs from that in the range $r < u < b$ (eq. (5-19b)). In other words, the integral is to be broken up into two parts so as to obtain

$$\begin{aligned} V &= GM \left(\int_{\infty}^b \frac{du}{u^2} + \int_b^r \frac{u^3 - a^3}{b^3 - a^3} \frac{du}{u^2} \right) \\ &= -GM \left(\frac{1}{b} + \frac{(b-r)(br(b+r) - 2a^3)}{2br(b^3 - a^3)} \right). \end{aligned} \quad (5-21)$$

In fig. 5-8(C), on the other hand, the field point P is located in the hollow interior of the

spherical shell and so the mass $M(r)$ in the interior of the Gaussian surface S is *zero*. Consequently,

$$I = 0 \quad (r < a). \quad (5-22)$$

In order to work out the potential, one again has to evaluate an integral of the form (5-20) where now the integral breaks up into *three* parts - one with u running from infinity down to b , one with u from b to a , and the last one with u from a to r . Of these the third contribution is zero by virtue of (5-22), and one then finds

$$V = -3G \frac{M(a+b)}{2(a^2 + ab + b^2)} \quad (r < a), \quad (5-23)$$

which is nothing but the expression (5-21) evaluated at $r = a$ (reason out why this should be so).

Problem 5-6

Intensity and potential in the interior of a homogeneous sphere.

Work out the gravitational intensity and gravitational potential due to a homogeneous sphere at an interior point.

Answer to Problem 5-6

A homogeneous sphere (with no hollow in its interior) of radius a (say) is a special case of a homogeneous hollow sphere, and the intensity and potential at various points can be obtained from corresponding results relating to the shell by the substitution $b \rightarrow a, a \rightarrow 0$. In particular, for an internal point, make use of the results in (5-19b) with the above substitution, to obtain

$$\mathbf{I} = -\frac{GM}{a^3} r \hat{r}, \quad (5-24a)$$

where r denotes the distance of the field point from the center of the sphere, and \hat{r} is the unit vector in the radial direction (i.e., $-\hat{r}$ is directed *towards* the center).

The potential at the interior point, similarly worked out from eq. (5-21), is

$$V(r) = -\frac{GM(3a^2 - r^2)}{2a^3}. \quad (5-24b)$$

.

Problem 5-7

How will equation (5-24b) be modified if the mass distribution within the sphere of radius a is spherically symmetric, but not uniform?

Answer to Problem 5-7

ANSWER:

$$V(r) = -\frac{G}{a} \left(M - a \int_a^r \frac{M(u)}{u^2} du \right), \quad (5-25)$$

where M stands for the total mass of the sphere and $M(r)$ is the mass contained within a spherical volume of radius r ($< a$)

One can now check that the limiting value of this potential for $r \rightarrow a$ is the same as the limit of the potential at an exterior point as the distance of that point from the center approaches a i.e., in other words, the potential is continuous across the surface of the sphere, as it should be.

Problem 5-8

A homogeneous solid sphere of radius a produces a gravitational intensity of magnitude g at its surface. At what two distances from its center will the intensity be $\frac{g}{4}$?

Answer to Problem 5-8

HINT: Of the two points, one will be in the interior of the sphere while the other will be located outside. In these two cases, use equations (5-24a), (5-18a) respectively to obtain the following results for the required distances from the origin: $d_1 = \frac{a}{4}$, $d_2 = 2a$. In this context see, sec. 5.4.1 and fig. 5-12.

5.3.3.3 Gravitational interaction of two spherical bodies

One interesting application of the result (5-18a) consists of working out the force of gravitational interaction between two bodies (say, A and B; see fig. 5-9) corresponding to non-overlapping spherical mass distributions. It might appear on the face of it that it would require the two mass distributions to be imagined as being made up of a large number of small volume elements, considering the forces between all possible pairs of such elements - one from each body (always taking care to focus on the force *on* the element in one particular body exerted *by* the element in the other body)-, and then summing up all these forces, which would reduce to integrations over the volumes of the two bodies. While this approach would work all right, it is, however, not as clever as the following one.

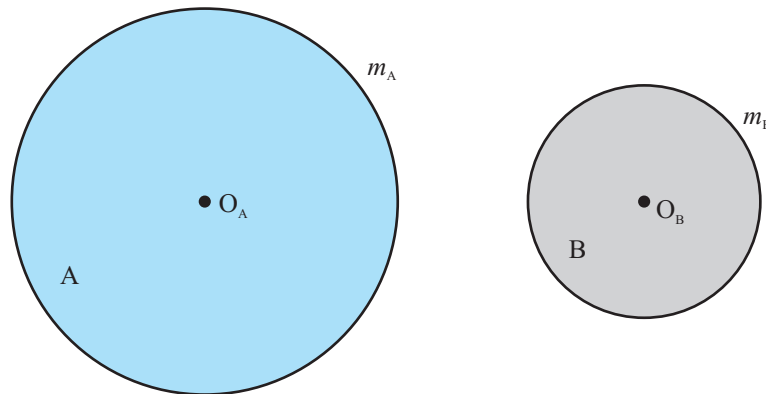


Figure 5-9: Gravitational interaction between two bodies A and B with non-overlapping spherical mass distributions; O_A and O_B are the centers of the two spherical bodies of mass m_A and m_B respectively; the force of gravitation between the two bodies is obtained by imagining each of these to be replaced with a single mass point (of mass equal to the corresponding body) located at its center.

We assume that the mass distributions are *rigid* in the sense that they do not get altered appreciably due to the gravitational pull of one on the other. The rest follows from the fact that the mass distributions are non-overlapping, i.e., each of the two bodies is external to the other.

Think of the force exerted by the body A on any small volume element of B. Since the

location of the latter is external to A, this force is the same as would result by assuming the entire mass of A (say m_A) to be concentrated at its center (O_A) (refer to eq. (5-18a)). Summing over the forces exerted by A over all such volume elements of B, one then concludes that the force exerted by A on B (call this F_{BA}) is the same as that exerted by the equivalent point mass m_A , placed at O_A , on the body B.

By the property (5-3b), this is equal and opposite to the force exerted *by* B *on* the point mass m_A located at O_A . But, with reference to the spherically symmetric mass distribution B, O_A is an external point, and hence, once again by the result (5-18a), this latter force is the same as that exerted by an equivalent point mass (say, m_B , the mass of B) located at the center O_B of B.

One thereby concludes that the gravitational force F_{BA} exerted by A on B is the same as that exerted by a point mass m_A , imagined to be located at O_A , on a second point mass m_B imagined to be located at O_B , i.e., is given by (see eq. (5-3a))

$$\mathbf{F}_{BA} = -\frac{Gm_A m_B}{r^3} \mathbf{r}, \quad (5-26)$$

where \mathbf{r} stands for the position vector of O_B relative to O_A .

The force exerted *by* B *on* A will, of course, be equal and opposite.

Problem 5-9

Obtain the gravitational intensity at a point close to the surface of a uniform disc of density ρ and thickness h , assuming that the point is far from the edge of the disc. Compare with the magnitude of intensity close to the surface of a homogeneous sphere of radius R , assuming that the density of the material of the sphere is the same as that of the disc.

Answer to Problem 5-9

HINT: Considering the point P close to the surface of the disc and the point P' symmetrically situated on the other side, as in fig. 5-10, the intensities at both points will be the same, say I (pointing towards and perpendicular to the disc surface), due to the symmetry of the two points

with respect to the disc (the intensity is the same, in an approximate sense, at all points close to the disc away from the edge; close to the edge, the field lines are curved, and the intensity varies sharply). Imagining a small cylindrical surface of area of cross section δS , the flux of intensity through it will be $-2I\delta S$ (reason out why; the intensity at any point on the curved surface of the cylinder is zero, again due to symmetry, while the flux through either of the end face is $-I\delta S$, where the minus sign appears due to the fact that the field points *inward* with reference to the cylindrical surface). According to the Gauss' principle, this must be equal to $-4\pi G\rho h\delta S$, which gives $I = 2\pi Gh\rho$.

Referring to the expression (5-18a) for the intensity at an external point of a homogeneous spherically symmetric mass distribution, the intensity at the surface of a homogeneous sphere of radius R is found to be $I' = \frac{4}{3}\pi G\rho R$, pointing towards the surface of the sphere, the density ρ being the same as for the disc material. This gives $\frac{I}{I'} = \frac{3h}{2R}$.

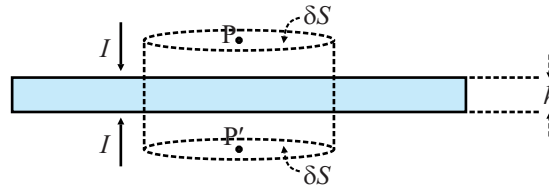


Figure 5-10: Intensity at a point P close to the surface of a disc (problem 5-9), where the point is assumed to be away from the edge of the disc; the intensity I , pointing towards the surface of the disc, is the same as that at a point P' on the other side, as can be seen from symmetry considerations; an imaginary cylindrical surface is shown dotted, with its curved surface perpendicular to the surface of the disc; the flux of intensity through the surface of the cylinder is $-2I\delta S$.

5.4 Earth's gravitational field: acceleration due to gravity

5.4.1 Earth's gravitational field

To a certain approximation, the mass distribution of the earth can be assumed to be a spherically symmetric one. For many practical purposes, the earth can even be described as a *homogeneous* spherical mass distribution of mass, say, M . The gravita-

tional intensity due to this mass distribution at an external point at a distance r from the center of the earth is, in accordance with eq. (5-18a),

$$\mathbf{I} = -g(r)\hat{r}, \quad (5-27a)$$

where \hat{r} is the unit vector along the radial direction (with the origin chosen at the center of the earth) at the point under consideration and

$$g(r) = \frac{GM}{r^2}, \quad (5-27b)$$

is termed the *acceleration due to gravity* at that point. It represents the magnitude of the acceleration that a particle placed at the point would acquire due to the earth's gravitational pull, the direction of the acceleration being toward the center of the earth.

If the point under consideration is on the surface of the earth or close to the surface, then one can use $r \approx R$, where R stands for the earth's radius. Thus, in an approximate sense, the acceleration due to gravity is the same at all points close to the surface of the earth, being given by

$$g = \frac{GM}{R^2}. \quad (5-28)$$

The symbol g is commonly used to denote this value of the acceleration due to gravity close to the earth's surface.

Problem 5-10

Using values $M = 5.97 \times 10^{24}$ kg, $R = 6.38 \times 10^6$ m, and G as in eq. (5-2), work out the value of g , the acceleration due to gravity at any point close to the surface of the earth.

Answer to Problem 5-10

ANSWER: Substituting in (5-28), $g = 9.8 \text{ m}\cdot\text{s}^{-2}$ (approx).

For a small region close to the surface of the earth at any place represented by, say, the point P in fig. 5-11, one can take, for any point, say, P' within this region, $r \approx R$, the radius of the earth, and $\hat{r} = \hat{n}$, where \hat{n} represents the unit vector along the vertically upward direction at P. In other words, the acceleration due to gravity may be taken to be *uniform* in magnitude *and* direction at all points throughout the region.

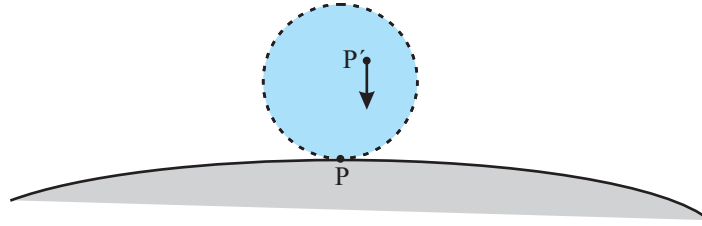


Figure 5-11: A point P on the surface of the earth and a small region close to it where the acceleration due to gravity may be assumed to be uniform, in the vertically downward direction; the gravitational force on a point mass at any point P' is independent of its location within the region.

Considering, finally, a point in the interior of the earth at a distance r ($< R$) and continuing to assume that the mass distribution of the earth is spherically symmetric, the acceleration due to gravity at the point is, in accordance with eq. (5-25)

$$g(r) = \frac{GM(r)}{r^2}, \quad (5-29a)$$

where $M(r)$ stands for the earth's mass included within a spherical volume of radius r . Assuming the earth to be a *homogeneous* spherical mass, this can be expressed as

$$g(r) = \frac{GM}{R^3} r, \quad (5-29b)$$

which tells us that, for an interior point, $g(r)$ is proportional to the distance from the center of the earth.

Incidentally, equations (5-27b) and (5-28) require only that the mass distribution be spherically symmetric (as does eq. (5-29a)), which need not be homogeneous. Fig. 5-12 depicts schematically the variation of acceleration due to gravity with distance from the

center of the earth, in which the linear part for $0 < r < R$ is obtained by assuming that the mass distribution is homogeneous.

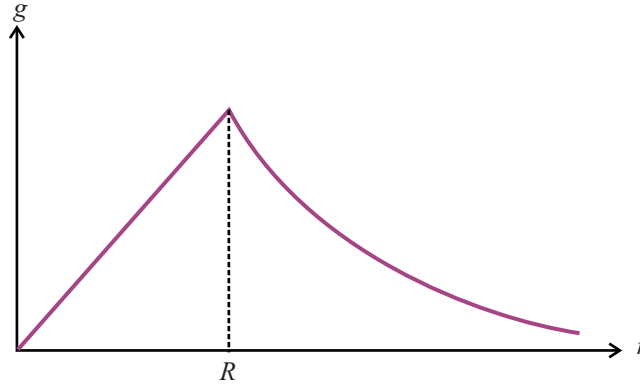


Figure 5-12: Graph depicting the variation of g , the acceleration due to gravity, with distance r from the center of the earth; with increasing r , g increases linearly in accordance with eq. (5-29b) (where it is assumed that the mass distribution is homogeneous) till r becomes equal to R , the earth's radius; for $r \geq R$, g decreases in accordance with eq. (5-27b).

The force of gravitation exerted by the earth is commonly referred to as *gravity*.

Problem 5-11

A satellite of mass m revolves with a period T in a circular orbit of radius R around a planet. If the magnitude of gravitational intensity at the surface of the planet is g , what is the radius of the planet, which may be assumed to be a homogeneous sphere?.

Answer to Problem 5-11

HINT: If the mass of the planet be M and its radius be a , then $g = \frac{GM}{a^2}$. The force of gravitation exerted by the planet on the satellite is $F = G \frac{Mm}{R^2}$. Since this provides for the centripetal acceleration, one has $G \frac{Mm}{R^2} = \frac{mv^2}{R}$, where v stands for the speed of the satellite in its circular orbit. Thus, $v = \sqrt{\frac{GM}{R}}$, and hence $T = \frac{2\pi R^{\frac{3}{2}}}{\sqrt{GM}}$. Eliminating M in favor of g one obtains $a = \frac{2\pi}{T} \left(\frac{R^3}{g} \right)^{\frac{1}{2}}$.

Problem 5-12

Geosynchronous orbit.

An artificial satellite, parked in the sky vertically above the equator appears to be stationary when seen from the earth. Find an expression for the height of the satellite above the surface of the earth?

Answer to Problem 5-12

HINT: The velocity v of the satellite, of mass m , in its circular orbit around the earth (termed a *geostationary* or a *geosynchronous* orbit) is given by $\frac{mv^2}{R+h} = G \frac{Mm}{(R+h)^2}$, where M stands for the mass of the earth, R for its radius, and h for the required height of the satellite above the earth's surface. In other words, $v^2 = \frac{gR^2}{R+h}$, where g stands for the acceleration due to gravity at the surface of the earth. The time period of the satellite ($\frac{2\pi(R+h)}{v}$) must be equal to the length of the day ($T = 24$ hr) for the satellite to appear stationary. This gives $h = \left(\frac{gT^2R^2}{4\pi^2}\right)^{\frac{1}{3}} - R$.

5.4.2 Center of gravity

Consider a rigid body held in any given orientation in space. It may be looked upon as a collection of particles rigidly attached with one another, where each of these particles is acted upon by the force of gravity. Assuming that the body is of a sufficiently small size compared to the earth, the acceleration due to gravity may be taken to be constant in magnitude and direction throughout the region occupied by the body. Thus, the forces on the particles making up the body constitute a system of like parallel forces, and the line of action of the resultant of these forces is a fixed line in the body under consideration, depending on its orientation. Considering now a different orientation of the body in a region of space where the acceleration of gravity has the same constant value as before, the line of action of the resultant force of gravity will be some other line fixed in the body. Considering in this manner the lines of action of the resultant force of gravity in all possible orientations of the body, all these lines will have a common point of intersection referred to as its *center of gravity*.

It is not difficult to see that, under the conditions mentioned above, the center of gravity is nothing but the center of mass of the body under consideration.

Problem 5-13

Establish the above statement, to the effect that the *center of gravity* of a rigid system in a uniform gravitational field is the same as its center of mass.

Answer to Problem 5-13

HINT: If, in any given orientation of the body, its constituent particles, of masses m_1, m_2, \dots, m_N , be at positions, say, $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N$, then the forces due to gravity acting through these points are, respectively, $m_1 g \hat{n}, m_2 g \hat{n}, \dots, m_N g \hat{n}$, where the acceleration due to gravity is g along the direction of the unit vector \hat{n} . The resultant of all these forces is $M g \hat{n}$, where $M = \sum_{i=1}^N m_i$ is the mass of the body, and the line of action of the resultant passes through the point $\mathbf{R} = \frac{\sum m_i \mathbf{r}_i}{\sum m_i}$, since the vector sum of the moments of all these forces about \mathbf{R} is zero (check this out, making use of the fact that the moment of the force of gravity on the i th particle is $m_i g \hat{n} \times (\mathbf{R} - \mathbf{r}_i)$). This establishes the statement since \mathbf{R} specifies the position of the center of mass of the body under consideration. It is a fixed point in the body since it can be checked that any rigid translation or rotation of the body results in the same translation or rotation of the point.

For a system of particles that may not necessarily be a rigid body, the resultant of the forces due to earth's gravitational attraction on the particles belonging to the system passes through its instantaneous center of mass, but the latter is, in general, not a point rigidly fixed in the system.

More generally, for a body of arbitrary size, the forces of gravity acting on the various particles of a body need not reduce to a single resultant since the acceleration due to gravity need not have a constant magnitude and direction throughout the region occupied by the body. The system of forces under consideration then reduces, in general, to a force and a couple.

Problem 5-14

Show that, if the earth be assumed to be a sphere with a spherically symmetric mass distribution, its gravitational pull on an external body, of arbitrary size and shape, reduces to a single force.

Answer to Problem 5-14

HINT: The forces on the particles making up the body constitute a concurrent system.

NOTE: This single resultant force, however, need not pass through the center of mass of the body under consideration. It is equivalent to an equal force through the center of mass and a couple in a plane containing the force and the center of mass.

5.4.3 The weight of a body

The force of gravity experienced by a body is commonly referred to as the *weight* of that body. Assuming the body to be made up of a large number of particles, each of these constituent particles is pulled by the force of gravity in a direction towards the center of the earth, and the weight of the body is the resultant of all these forces.

If the size of the body is assumed to be sufficiently small so that the acceleration due to gravity may be taken to be constant in magnitude (g) and direction throughout the region of space occupied by it, the weight of the body can be expressed in the form

$$W = mg, \quad (5-30)$$

where m stands for the mass of the body. The weight will act through the center of gravity of the body, and its direction will be that of the acceleration due to gravity (see sec. 5.4.2).

If, in particular, the body is located on the surface of the earth, then the weight will act in the vertically downward direction.

5.4.3.1 Weight as a force of reaction: weightlessness

A person standing on the floor of a room feels her weight in the form of the *force of reaction* exerted by the floor on her body. In order that her body be in equilibrium, this

force of reaction (say, R) has to be equal to the force of gravity on her body,

$$R = W, \quad (5-31)$$

and, moreover, has to act in the vertically upward direction, its line of action being through her center of gravity (if she is not to experience a torque tending to make her fall).

Suppose, now, that the same person is located in an elevator which is ascending vertically with a uniform acceleration f . In the frame of reference of the elevator, a pseudo force mf acts on the body of the person, the direction of this pseudo force being vertically downward. This will then add to the force of gravity acting on her, and the reaction force exerted by the floor of the body of the person will be

$$R = W + mf, \quad (5-32a)$$

which she will now feel as her weight, i.e., the weight in the frame of reference of the elevator (see fig. 5-13). In a similar manner, if the elevator *descends* with an acceleration f (incidentally, it is the *direction* of acceleration that matters in either case, regardless of the direction of motion; for instance, if the elevator ascends with a deceleration f , then formula (5-32b) applies), the reaction force, and hence the weight in the frame of the elevator will get reduced to

$$R = W - mf. \quad (5-32b)$$

In particular, if the elevator descends *freely* under gravity, i.e., the downward acceleration of the elevator be g , the acceleration due to gravity, then eq. (5-32b) tells us that the person will not feel any reaction force at all, i.e., in other words, she will feel *weightless*.

The same phenomenon of weightlessness is experienced by an astronaut in an artificial satellite. Assuming that the satellite is at a location where the acceleration due to gravity is g , and that the only force on it is the one due to gravity, its acceleration relative to

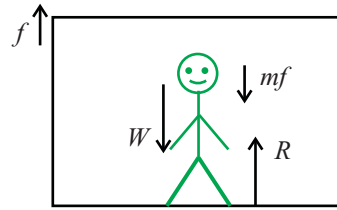


Figure 5-13: Weight of a body in an accelerating lift, represented by the force of reaction exerted on it by the floor of the lift; if the lift has an upward acceleration f then, in the frame of reference of the lift, the forces acting on the body are the gravitational pull W and the inertial force mf , both acting downwards, and the reaction force R acting upwards; the weight of the body, measured in terms of the reaction force, exceeds W ; for a lift accelerating downward, the weight will be less than W .

the frame of reference of the earth (an approximately inertial frame) will be g , and hence in the frame of reference of the satellite, a pseudo force mg will act on the body of the astronaut in a direction opposite to that of the acceleration due to gravity.

Thus, the resultant force on her body in the frame of reference of the satellite will be zero and she will be able to float around within it. In particular, there will be no force of reaction even when she stands on the floor of the satellite.

Similar considerations apply to space flight as well. Indeed, the force of gravitation acting on the satellite as also on the astronaut need not be that due to the earth alone, and the gravitational pull by other heavenly bodies may also be relevant in determining the effective gravitational force on a body which can be written in the form mg , where m is the mass of the body and g is the effective 'acceleration due to gravity'.

What is important is to note that the same value of g will account for the acceleration of all bodies at the given location in space regardless of their masses, a fact that follows from Newton's law of gravitation as expressed in eq. (5-6b). Assuming that the spaceship or the artificial satellite experiences only this force of gravitation at the location under consideration (in which case it is said to be 'in free fall', in analogy to an elevator falling under gravity with an acceleration g), its acceleration in an inertial frame will be g and hence, in the frame of the spaceship a pseudo force will act on the astronaut, exactly canceling the force on her body due to the gravitational field at the location of the spaceship.

5.4.3.2 Weight reduction due to earth's rotation

Fig. 5-14 depicts a point mass m at a point P near the surface of the earth, where the latitude of P is, say, λ . The diurnal rotation of the earth takes place with an angular velocity ω ($= \frac{2\pi}{24 \times 3600} \text{ rad} \cdot \text{s}^{-1}$) about its axis ON where O represents the center of the earth's sphere and N the north pole. The gravitational pull of the earth on the point mass is mg acting along PO, g being the 'true' acceleration due to gravity.

However, for an earth-bound observer, this is not the only force acting on the mass. A frame of reference fixed rigidly to the earth rotates with it with an angular velocity ω , and hence is a non-inertial frame, in which the point mass (assumed to be at rest) experiences a centrifugal force (see sec. 3.21) of magnitude $m\omega^2\rho$, where $\rho = R\sin\theta$ (R = earth's radius; $\theta = \frac{\pi}{2} - \lambda$ is referred to as the co-latitude of the point P).

The resultant force on the point mass at P acts along the line PO' and is of magnitude mg' , where

$$g' \approx g\left(1 - \frac{\omega^2 R}{g} \sin^2 \theta\right) = g\left(1 - \frac{\omega^2 R}{g} \cos^2 \lambda\right). \quad (5-33)$$

Problem 5-15

Check eq. (5-33) out.

Answer to Problem 5-15

HINT: Make use of the formula for the resultant of two given forces (or, equivalently, for the resultant of two given vectors), and of the approximation $\omega^2 R \ll g$, checking out the validity of this approximation.

One thus finds that $g' < g$, where g' represents the 'effective' acceleration due to gravity at a latitude λ . A plumb line suspended at P will hang along the line QPO' (see figure 5-14), the 'effective' vertical direction at P, instead of along PPO, the radial direction toward the earth's center.

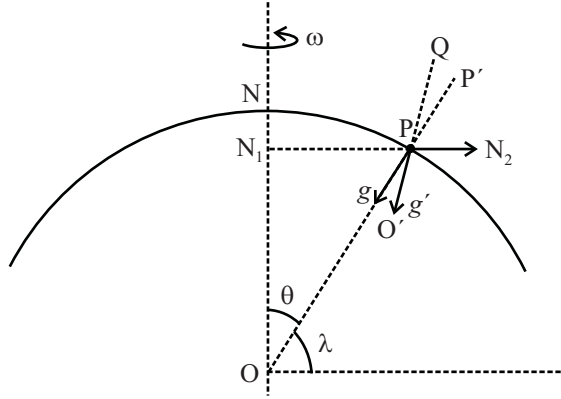


Figure 5-14: Reduction of weight due to the diurnal rotation of earth; the diurnal rotation occurs about the axis ON with angular velocity ω ; P is a point near the surface of the earth with co-latitude θ ; the acceleration due to the earth's pull on a point mass m at P is g along PO , where O is the center of the earth, assumed to be spherical in shape; the centrifugal force acts along PN_2 , where N_1PN_2 is perpendicular to the axis ON ; the resultant force on the point mass is mg' along PO' ; OP' is the geographical vertical direction at P , while a plumb line suspended at P hangs along QPO' ; the various directed line segments shown are not to scale.

The reduction in the value of g , the acceleration due to gravity implies a corresponding reduction in the weight $W = mg$ of the particle as well, the fractional reduction in weight being $\frac{\omega^2 R}{g} \cos^2 \lambda$. This reduction being proportional to the squared cosine of the latitude, attains its maximum value at the equator ($\lambda = 0$), and is zero at the poles ($\lambda = \pm \frac{\pi}{2}$).

5.4.4 Escape velocity

Consider a particle of mass m located at a point P at a distance r_0 from the center of the earth where, for the sake of concreteness, we take $r_0 \geq R$, the earth's radius. Let the speed of the particle at this point be v_0 . Then, using the expression (5-18c), the total energy of the particle is seen to be

$$E = \frac{1}{2}mv_0^2 - \frac{GMm}{r_0}. \quad (5-34)$$

Suppose that, by virtue of the velocity at the point P , the particle eventually escapes from the earth's gravitational field (the field of force exerted on it due to gravity) and moves up to an infinitely large distance, being finally left with a speed, say, v' . Noting from eq. (5-18c) that, at an infinitely large distance from the center of the earth, the

potential energy of the particle is zero, its total final energy works out to $\frac{1}{2}mv'^2$. Making use, then, of the principle of conservation of energy (recall that the gravitational force field is a conservative one), one obtains

$$\frac{1}{2}mv_0^2 - \frac{GMm}{r_0} = \frac{1}{2}mv'^2. \quad (5-35)$$

Note that this equation has been arrived at on the assumption that the particle, starting from the point P with a velocity v_0 , moves up to an infinitely large distance, where its speed has been denoted by v' . An alternative possibility is that the particle, starting with the above initial condition, *fails* to reach up to an infinite distance. These two possibilities are distinguished by referring to the motion of the particle as an *unbounded* or a *bounded* one respectively.

Since the right hand side of equation (5-35) is necessarily a non-negative expression, one straightaway arrives at the conclusion that, for the motion to be an unbounded one, a necessary condition is

$$\frac{1}{2}mv_0^2 - \frac{GMm}{r_0} \geq 0, \quad (5-36a)$$

or, in other words, for the particle starting from the point P at a distance r_0 from the center of the earth to reach up to an infinite distance, its speed v_0 has to satisfy

$$v_0 \geq \sqrt{\frac{2GM}{r_0}}. \quad (5-36b)$$

Conversely, the condition for *bounded* motion of a particle in an attractive inverse square central field of force (i.e., a field where the force varies inversely as the square of the distance from a fixed center and is always directed towards the latter) is that its total energy, which remains constant during its motion along any particular trajectory, is negative.

The minimum speed ($\sqrt{\frac{2GM}{r_0}}$) that the particle must possess at P so as to be able to

escape from the earth's gravitational field, i.e., to move up to an infinitely large distance, is referred to as the *escape velocity* for the point P (recall that the term 'velocity' is sometimes used when the more precise term to use would be 'speed').

If, in particular, the point P is close to the surface of the earth, then one has

$$v_{\text{escape}} = \sqrt{\frac{2GM}{R}} = \sqrt{2gR}, \quad (5-37)$$

where R stands for the radius of the earth, and g denotes the acceleration due to gravity at the earth's surface (refer to eq. (5-28)).

Problem 5-16

Using $g = 9.81 \text{ m}\cdot\text{s}^{-2}$, and $R = 6.38 \times 10^6 \text{ m}$, work out the value of the escape velocity from the surface of the earth.

Answer to Problem 5-16

ANSWER: $v_{\text{escape}} = 11.19 \text{ km}\cdot\text{s}^{-1}$.

5.5 The motion of planetary bodies

5.5.1 Introduction

The solar system consists of the sun as the principal gravitating body, with all the planets moving around in the gravitational field of the sun. The planets exert their own gravitational pull on one another, but these can be ignored as negligible to start with. In other words, each planet can be assumed to move in the gravitational field of the sun independently of other planets in the system. The motion of satellites is also determined, in the main, by the solar gravitational field, with the gravitational fields of the respective planets also having a relevant role.

In describing the motion of a planet around the sun, one can consider a frame of reference in which the sun is at rest since that frame is, to a good degree of approximation,

an inertial one (in reality, however, the center of mass of sun and the planet defines an inertial frame to a better degree of approximation; the solar mass being large, one may as well consider this as the solar frame). Assuming that the sun can, moreover, be described as a spherically symmetric mass distribution, its gravitational field at external points is equivalent to that of a point particle of mass M (say), the solar mass, located at the center of the sun. A convenient description of the planet's motion around the sun is obtained by replacing it with a point mass as well, which gives the motion of the center of mass of the planet.

In other words, one is led to consider the motion of a point particle of mass, say, m (the planetary mass), in the gravitational field of a particle of mass M fixed at a point, which one can choose as the origin.

One can then write down the equation of motion of the planet (now represented by the point mass m) in the gravitational field of the sun (fixed particle of mass M) and solve it, subject to appropriate *initial conditions* to obtain the *trajectory* of the planet around the sun. This trajectory is, in general, found to be an *elliptic* one. In arriving at a successful explanation of such elliptic planetary orbits, Newtonian mechanics established itself as a major foundation-stone of physics.

5.5.2 The equation of motion and the nature of trajectories

As mentioned above, the frame of reference in which the sun is at rest is, to a good degree of approximation, an inertial frame, and the equation of motion of the planet (refer to eq. (3-50a)), with the force on the planet described by Newton's law of gravitation, can be written as

$$m \frac{d^2 \mathbf{r}}{dt^2} = - \frac{GMm}{r^2} \hat{r}, \quad (5-38)$$

where \mathbf{r} stands for the instantaneous position vector of the planet with respect to the sun (see fig. 5-15, recall that the sun and the planet have been reduced to point masses in our description), and \hat{r} is the unit vector along \mathbf{r} .

To be more precise, however, the mass m in the left hand side of the above equation is to be replaced with $\mu \equiv \frac{mM}{M+m}$, the *reduced mass* of the sun and the planet. However, the solar mass M being large compared to the mass m of the planet, one has $\mu \approx m$.

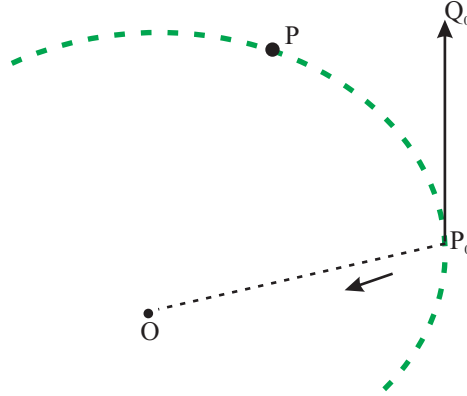


Figure 5-15: Depicting a part of the trajectory (schematic) of a planet imagined to be represented by a point mass in an inverse square field around the origin O ; P_0 is the initial position of the planet, when the force on it is along P_0O ; the initial velocity is directed along P_0Q_0 ; the trajectory is a planar one, confined to the plane containing O and the line P_0Q_0 ; for any other position P on the trajectory, the force is directed along PO .

The equation (5-38) is referred to as the equation of motion of a particle in an *inverse square* field of force, and is of considerable relevance in describing electronic orbits around the nucleus in an atom (see chapters 16 and 18), though the concept of an orbit inside an atom is only of limited relevance.

On the face of it, eq. (5-38) describes a three dimensional motion in space. On analyzing the equation, however, it is found that *the motion remains confined to a plane* determined by the *initial conditions*. More concretely, if at any given instant t_0 , the particle (representing the planet in the present context) is located at P_0 (fig. 5-15), and its velocity is directed along, say, P_0Q_0 , then the motion remains confined to the plane containing OP_0 and P_0Q_0 at all times.

While remaining confined to a plane, the motion of the particle representing the planet can, as I have mentioned above, possibly be either of two types - *bounded* and *unbounded*. In a bounded motion, the distance of the particle from the origin O remains

finite at all times while, in an unbounded one, the distance increases beyond all finite values at sufficiently large times.

Eq. (5-36b) gives the condition for unbounded motion, where the symbols are to be re-interpreted in the present context in an obvious manner. The condition for bounded motion can then be stated in the form

$$v_0^2 < \sqrt{\frac{2GM}{r_0}}. \quad (5-39)$$

The next interesting fact that comes out of looking at the general nature of the solution to the equation of motion (5-38) is that, for any given initial condition, whether satisfying eq. (5-36b) or (5-39), the trajectory of the particle is a curve belonging to the class of *conic sections*.

In the case of unbounded motion, the conic section can be either a parabola, or a branch of a hyperbola while it may, under special initial conditions, even reduce to a straight line extending to an infinitely large distance from the origin.

The motion of a planet around the sun, on the other hand, is a *bounded* one. A conic section that remains within a bounded region in a plane is, in general, an *ellipse*. A solution of the equation of motion, eq. (5-38), moreover, shows that one of the two foci of the ellipse lies at the point O. Thus, in other words, *the trajectory of a planet around the sun is an elliptic one, with the sun located at one of the two foci of the ellipse*.

5.5.3 Kepler's laws

A large body of data accumulated through observations on planetary motion was summarized by Kepler in the form of three laws. I include below a statement and a brief explanation of these three. All these three laws of Kepler follow from a solution of the basic equation, eq. (5-38), which can therefore be taken as a validation of the fundamental principles of Newtonian mechanics by empirical observations. As you will presently see, the first of the three laws of Kepler relates to a conclusion from the equation of

motion we have already come across in the last section. The remaining two are also similarly derived from the equation of motion.

What is more, the relation between the equation of motion (5-38) and the three laws of Kepler can also be seen to work backwards, i.e., *starting* from the three laws, one finds that these can be explained only on the basis of an equation of motion of the form (5-38). In this sense, the three laws of Kepler stated below can be looked upon as enumerating the *basic* principles of planetary motion. Here are the three laws:

1. A planet describes an elliptical trajectory around the sun, with the sun being located at one of the foci of the ellipse.
2. As the planet moves along its elliptic trajectory, its *areal velocity* with reference to the sun remains constant.
3. The square of the time period of complete revolution of a planet around the sun is proportional to the cube of the length of the semi-major axis of its elliptic orbit.

Since we have already been acquainted with the first of these laws, I will briefly explain below the meaning of the second and the third laws. But before that I want to state that these laws of Kepler, while formulated in the context of planetary motion, apply to the motion of any particle moving in an attractive inverse square field of force, i.e., one where the force on the particle at any given position is of the form $-m\frac{\gamma}{r^3}\mathbf{r}$ ($\gamma > 0$), where \mathbf{r} denotes the instantaneous position vector of the particle relative to the center of attraction for the attractive inverse square force field. As I mentioned in sec. 5.5.2, the motion of a planet under the gravitational attraction of the sun can indeed be reduced, in an approximate sense, to that of a particle in such an attractive inverse square field, with $\gamma = GM$.

Fig. 5-16 shows an elliptic orbit of a particle in an inverse square attractive field of force, with the center of attraction O (corresponding to the sun in the solar system) at one focus of the ellipse, where P and P' denote two successive positions of the particle at time instants, say t and $t + \delta t$, δt being a small interval of time. The area bounded by the lines OP, OP', and the arc PP' denotes the area swept by the radial line OP in time

δt . If this area be denoted by δA , then the *areal velocity* of P with reference to the focus O is defined as $\frac{\delta A}{\delta t}$ in the limit of infinitesimally small δt . Kepler's second law then states that this rate at which the area is described is the *same* for all positions P of the particle on the ellipse.

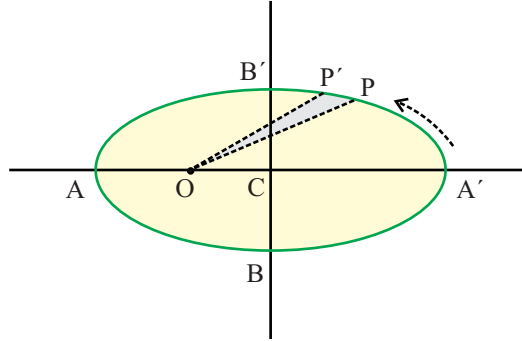


Figure 5-16: The elliptic orbit of a point mass in an attractive inverse square field of force, showing major and minor axes AA' and BB'; C is the center of the ellipse and O is the center of attraction, located at one focus of the ellipse; P and P' are two positions of the particle at times t and $t + \delta t$; the shaded area then represents the area swept by the radial line OP in time δt , from which one obtains the areal velocity.

Finally, the third law of Kepler expresses a relation of proportionality between the time period (say, T) of revolution of the particle around the elliptic orbit and the major axis (say, a) of the ellipse. Thus, considering two possible elliptic orbits, corresponding to two different sets of initial conditions, in the same inverse square field of force, if T_1 and T_2 be the time periods on these orbits and a_1 , a_2 be the semi-major axes of the orbits, then

$$\frac{T_1^2}{T_2^2} = \frac{a_1^3}{a_2^3}. \quad (5-40a)$$

For a planet in the solar system, the semi-major axis is half the sum of the smallest and largest distances from the sun in one complete revolution, and the above formula, applied to any two planets, relates the ratio of their time periods (respective solar years) to the ratio of their lengths of semi-major axes.

An alternative way to express the above formula is to state that the ratio $\frac{T^2}{a^3}$ for any

given elliptic orbit (T = time period, a = length of semi-major axis) is independent of the initial conditions of the orbit or of the mass of the particle, being determined solely by the constant γ for the inverse square field. Indeed, a solution of the equation of motion (eq. (5-38)) gives the result

$$\frac{T^2}{a^3} = \frac{4\pi^2}{\gamma}, \quad (5-40b)$$

where, recall that, in the case of a planetary orbit, $\gamma = GM$.

Problem 5-17

A particle describes the elliptic orbit shown in fig. 5-16 in an inverse square attractive field, with O as the center of attraction. Its velocities at the points A, A' (points at minimum and maximum distances from O) are, respectively, v, v' . If $OA = l$, find the distance OA' , and hence the eccentricity of the ellipse. What is the time period of revolution in the ellipse?

Answer to Problem 5-17

HINT: Since the direction of motion at either of the points A, A' is perpendicular to the axis AA', the constancy of the areal velocity implies $\frac{1}{2}lv = \frac{1}{2}l'v'$, where l' is the required distance of the farthest point A' from the center of attraction O. This gives $l' = \frac{lv}{v'}$. If the eccentricity of the ellipse be e then, according to the geometry of the ellipse, $\frac{l'}{l} = \frac{1+e}{1-e}$, i.e. $e = \frac{v-v'}{v+v'}$. The areal velocity being $\frac{1}{2}lv$, and the area of the ellipse being $A = \pi ab = \frac{\pi l^2 \sqrt{1-e^2}}{(1-e)^2}$ (check this out; a, b stand for the semi-major and the semi-minor axes of the ellipse), the time period of revolution is obtained by substituting for A in the expression $T = \frac{2A}{lv}$.

5.5.4 Circular orbits in an inverse square field

A special case of a bounded orbit of a particle in an attractive inverse square field of force corresponds to a *circular* orbit, a circle being a special instance of an ellipse where the eccentricity of the ellipse is zero, i.e., where the major and the minor axes are equal. The two foci of the ellipse are then coincident and correspond to the center of

the circle. Moreover, the constancy of the areal velocity about the center implies that particle actually executes a *uniform* circular motion in this special case.

Problem 5-18

Show that the constancy of the areal velocity in a circular motion does indeed imply a constant speed.

Answer to Problem 5-18

HINT: If the instantaneous speed of a particle moving on a circular orbit of radius a is v then, in a small time interval δt , the particle moves through a distance $v\delta t$ in a direction perpendicular to the radius vector stretching from the center to the instantaneous position of the particle, which means that the area described by the radius vector is $\frac{1}{2}av\delta t$. Hence the instantaneous areal velocity is $\frac{1}{2}av$.

Recall by referring to section 3.19.4, that a particle with speed v executing a circular motion on a circle of radius a possesses a *centripetal* acceleration, i.e., one directed radially inward, of magnitude $\frac{v^2}{r}$. This means that there has to be a radially inward force $\frac{mv^2}{r}$ acting on the particle so as to *make it move* along the circular path.

Suppose, then, that at any given instant $t = t_0$, the velocity \mathbf{v}_0 of a particle moving under an attractive central field of force is along a direction perpendicular to the radius vector \mathbf{r}_0 from the center of the force to its instantaneous position and that, further, the condition

$$\frac{mv_0^2}{r_0} = \frac{m\gamma}{r_0^2}, \quad (5-41)$$

is satisfied, where $v_0 = |\mathbf{v}_0|$, $r_0 = |\mathbf{r}_0|$, and the force field is given by $\mathbf{F}(\mathbf{r}) = -\frac{m\gamma}{r^3}\mathbf{r}$, all position vectors being relative to the center of the force.

Under *this* initial condition, the particle will continue to move along a circular orbit of

radius $a = r_0$ with uniform speed

$$v = v_0 = \sqrt{\frac{\gamma}{a}}. \quad (5-42)$$

Referring to condition (5-36b) (with $GM = \gamma$), one can check that eq. (5-41) implies a bounded orbit, as it should.

The time period T for uniform motion with speed v in a circular orbit of radius a is given by $T = \frac{2\pi a}{v}$. Using eq. (5-42) one gets

$$T = \frac{2\pi}{\sqrt{\gamma}} a^{\frac{3}{2}}. \quad (5-43)$$

As expected, this conforms to Kepler's third law, and is equivalent to eq. (5-40b) in the special case of an elliptic orbit reducing to a circular one.

5.5.5 Tidal force

Fig. 5-17(A) shows three bodies A, B, C (which we consider to be point masses for the sake of simplicity) of masses M , m , and m' respectively. Assuming that the three are free from the influence of all other bodies, the equation of motion of C in the gravitational field of A and B, as seen from an inertial frame S, can be written as

$$m'\ddot{\mathbf{r}}_C = \mathbf{F}_C \text{ (say)} = -Gm' \left(M \frac{\mathbf{r}_C - \mathbf{r}_A}{|\mathbf{r}_C - \mathbf{r}_A|^3} + m \frac{\mathbf{r}_C - \mathbf{r}_B}{|\mathbf{r}_C - \mathbf{r}_B|^3} \right) = \mathbf{F}_{CA} + \mathbf{F}_{CB} \text{ (say)}, \quad (5-44)$$

where \mathbf{r}_A , \mathbf{r}_B , \mathbf{r}_C , stand for the position vectors of the three particles in the frame of reference under consideration relative to any chosen origin.

Suppose now that one chooses to describe the motion of C from a frame S_B fixed to B. Let us assume for the sake of simplicity that the gravitational force of C on B can be ignored as being small compared to that of A on B. The frame S_B attached to B then possesses an acceleration $\mathbf{f}_B = -GM \frac{\mathbf{r}_B - \mathbf{r}_A}{|\mathbf{r}_B - \mathbf{r}_A|^3}$ relative to S. The equation of motion of C in

S_B will then be of the form

$$m'\ddot{\mathbf{r}}_{CB} = \mathbf{F}_C + \mathbf{G}_C = \mathbf{F}_{CA} + \mathbf{F}_{CB} + \mathbf{G}_C, \quad (5-45)$$

where \mathbf{G}_C is the inertial force on C appearing in the frame S_B , and where \mathbf{r}_{CB} stands for the position vector of C in the frame S_B in which the origin is chosen to be at B for the sake of simplicity.

The position vectors appearing in the expressions for the forces get transformed in going over from one frame to another, but the forces themselves remain unchanged since these depends on the *relative* separations between the particles involved (refer to eq. (3-46)).

Recall from sec. 3.10.3 that the inertial force \mathbf{G}_C in S_B is of the form $-m'\mathbf{f}_B$, and one can then rearrange terms in eq. (5-45) as

$$m'\ddot{\mathbf{r}}_{CB} = \mathbf{F}_{CB} + \mathbf{F}_T, \quad (5-46a)$$

where

$$\mathbf{F}_T = -GMm' \left(\frac{\mathbf{r}_{CA}}{|\mathbf{r}_{CA}|^3} - \frac{\mathbf{r}_{BA}}{|\mathbf{r}_{BA}|^3} \right), \quad (5-46b)$$

is referred to as the *tidal force* on C as seen from B. The symbols \mathbf{r}_{CA} and \mathbf{r}_{BA} in the above equation stand for the separations between C and A and B and A respectively (check equations (5-46a), (5-46b) out.)

Strictly speaking, the mass m' in the left hand sides of equations (5-45), (5-46a) should be replaced with $\frac{mm'}{m+m'}$, a quantity referred to as the *reduced mass* of B and C. This approximates to m' if m' is small compared to m ($m' \ll m$). I do not enter into the details here, assuming for the sake of simplicity that the approximation $m' \ll m$ applies to the system under consideration.

The tidal force is thus of a simple interpretation: it is the difference of the gravitational pulls on C and B exerted by A. Put differently, the pseudo force on C in the frame S_B cancels a part of the gravitational pull on C exerted by A, and what remains after the cancellation is the tidal force.

Eq. (5-46a) tells us that the motion of C as seen from B is governed by the forces F_{CB} , the gravitational pull on C exerted by B, and F_T , the tidal force on C. For numerous problems of interest, this is a convenient way of describing the motion of C, since the tidal force in these cases turns out to be only a small correction term over the gravitational pull F_{CB} on C.

This happens, for instance, when the separation r_{CB} is small in magnitude compared to r_{BA} . Let us assume, for the sake of simplicity, that A, B, and C all lie on a straight line (fig. 5-17(A)), and let the distances between A and B and between B and C be D and d respectively, where $d \ll D$. One then has

$$F_{CB} = -\frac{Gmm'}{d^2}, \quad F_T = -GMm' \left(\frac{1}{(D+d)^2} - \frac{1}{D^2} \right) \approx GMm' \frac{2d}{D^3}, \quad (5-47)$$

The ratio of the magnitudes of the tidal force and the gravitational pull on C due to B is given by

$$\frac{|F_T|}{|F_{CB}|} = \frac{2M}{m} \left(\frac{d}{D} \right)^3. \quad (5-48)$$

The tidal force is responsible to a large extent for the formation of tides in the oceans, where A, B, and C correspond to the moon, the geosphere, and a mass of oceanic water on the earth's surface respectively. In this instance, the earth being an extended body, the second term on the right hand side in eq. (5-46b) is to be interpreted as the gravitational pull (with a negative sign) exerted by the moon (a spherical body) on a particle of mass m' imagined to be placed at the center of the earth. In that case, D and d would correspond the earth-moon distance and the earth's radius respectively.

5.5.6 The orbit of the moon

It was Newton who first gave quantitative results relating to the moon's orbit around the earth and its time period of revolution by making use of the laws of gravitation. However, his results were only a first approximation and a great many corrections have since been incorporated for an accurate determination of the orbit of the moon, the determination of these corrections constituting, in general, a complex problem.

The orbit of the moon can be looked at either from an inertial frame, say a frame in which the sun is at rest, or from an earth-bound frame. Let us first consider the motion of the moon as seen from the earth. On the face of it, it would appear that the pull exerted by the sun on the moon should have an important role to play in determining this motion. In fact, however, a large part of this gravitational pull is canceled by the pseudo force acting on the moon in the earth-bound frame, leaving only the *tidal force* as the residue. Thus one has to consider, first, the earth's pull on the moon and then this tidal force. Of these two, the earth's pull constitutes by far the dominant factor and the tidal force can be considered to be only a small correction. The ratio in eq. (5-48) in this instance turns out to be of the order of five parts in a thousand.

Thus, to a first approximation, the orbit of the moon as seen from the earth is simply explained in terms of the earth's gravitational pull on the moon and is very much similar to the orbital motion of the earth itself around the sun. But then one has to consider the effect of the tidal force, which is a periodically varying one because of the fact that the relative orientation of the vectors \mathbf{r}_{CA} and \mathbf{r}_{BA} (here A, B, C correspond to the sun, the earth, and the moon respectively) vary with time, with a periodicity approximately equal to the lunar month. When this periodically varying tidal force is taken into consideration as a correction over the gravitational pull of the earth, one gets an improved description of the lunar motion as seen from the earth.

On the other hand, as seen from an inertial frame in which, say, the sun is at rest, the moon executes an orbit *around the sun*. In reality, both the earth and the moon

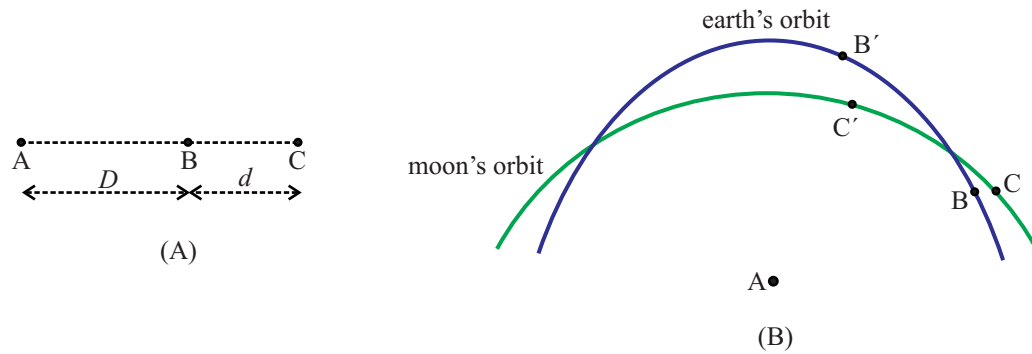


Figure 5-17: (A) Illustrating the tidal force; A, B, C represent three bodies, of which the motion of C in the frame attached to B is affected by the tidal force on C; the latter is the *differential* gravitational pull on C and on B exerted by A; the three bodies have been taken to lie on a single straight line for the sake of illustration; in the case of oceanic tides, A, B, and C correspond to the moon, the earth's sphere, and a body of oceanic water on the surface of the earth's sphere; (B) depicting the variation with time of the relative orientation of the vectors extending from the sun to the earth and from the sun to the moon respectively; A, B, C correspond to the sun, the earth, and the moon; B' and C' are the positions of the earth and the moon at a different point of time; as a result, the tidal force on the moon in an earth-bound frame varies periodically; in this case, the distances D and d in fig. (A) correspond to the sun-earth and the earth-moon distances respectively; parts of the orbits of the earth and the moon are depicted schematically showing how the latter crosses into and out from the former; both are everywhere convex as seen from a point exterior to the two orbits.

follow very similar orbits around the sun, with the moon's orbit, which is additionally influenced by the earth's pull on the moon, having a number of subtle and complex features. For instance, while the earth's orbit around the sun is very nearly an ellipse (which is, moreover, close to being a circle) the moon's orbit *crosses* the earth's orbit twice during a lunar month. Interestingly, in spite of the lunar orbit crossing periodically into and out of the earth's orbit, the former is everywhere a *convex* curve like the latter (or, more precisely, the boundary curve of a convex set, see fig. 5-17(B); the lunar orbit looks approximately a polygon with thirteen sides, having rounded corners).

The motion of the earth around the sun can be described up to a good degree of approximation as a *two-body* problem in mechanics where the sun's gravitational attraction on the earth explains quite accurately the earth's orbit. The problem of an equally accurate description of the lunar orbit, on the other hand is much more non-trivial since it is, in effect, a *three-body* problem in mechanics involving the sun, the earth, and the moon. The three body problem or, more generally, the N -body problem in mechanics involving $N(\geq 3)$ gravitating bodies is one of the great *unsolved* problems in theoretical

physics, so far as exact solutions are concerned. However, the sun-earth-moon problem has a number of simplifying features that allows for a calculation of the lunar orbit in approximate terms, where such calculations have been carried out to a very high degree of accuracy.

5.5.7 The motion of a projectile

Analogous to the motion of the earth along an elliptic orbit around the sun, one may consider the motion of a body in the gravitational field of the earth. Assuming the earth's attraction to be the only force acting on the body, and imagining the body to be represented by a point mass, one is once again led to consider the motion of a particle in an attractive inverse square central field of force. As we have seen in sec. 5.5.2, the trajectory described in such a motion is a conic section. For instance, in the case of a bounded motion, the trajectory is, in general, an ellipse with the earth's center as one of the two foci. On the other hand, the motion of a projectile under earth's gravity is known to be a parabola (see problem with hint in sec. 3.7). There is, however, no conflict between these two statements.

The parabolic trajectory of a projectile is obtained on the assumption that the earth's gravitational pull on the particle under consideration is *constant* in magnitude and direction, which is an approximation valid within any small region of space. If the trajectory of the particle is considered over a larger region of space then the variation of the gravitational force exerted on the particle by the earth has to be taken into consideration, which then implies an elliptic trajectory. Put differently, the parabolic trajectory of a projectile is an approximate description, valid over a small region of space, of the trajectory, which is actually an elliptic one.

The description of the motion of the projectile given above is with respect to an earth-bound frame of reference. In this frame, there are actually *three* forces acting on the projectile: the pull exerted by the earth, that exerted by the sun, and the *pseudo-force* arising due to the motion of the earth in the gravitational field of the sun (we assume that a frame of reference fixed with respect of the sun is an inertial one). The second

and third of these cancel out among themselves (with only a small tidal force remaining as the residue, which one can ignore; refer to sec. 5.5.5), leaving only the pull of the earth to account for the motion of the projectile. In addition, if the frame be assumed to share the diurnal rotation of the earth, then an additional pseudo-force due to that rotation is to be taken into consideration which, however, has been ignored here.

5.6 Gravitation: a broader view

In trying to give you the broader view of gravitation, I feel way out of my depth since it involves a breathtakingly wide range of fundamental issues in physics. Indeed, gravitation, in a sense, constitutes *the* central problem in theoretical physics of recent times. Here is what I understand the reason to be.

On the face of it, gravitation seems to fit in with the Newtonian view of physics where it is looked at as just a basic force of nature analogous to the Coulomb force between charges, or the magnetic force between currents. From this point of view, Newton's law of gravitation has been remarkably successful in explaining an astoundingly large range of observed facts, including the motions of celestial bodies. But the story has its twists.

The electrical and magnetic forces are known to be two aspects of a single basic phenomenon, namely, the interaction between particles mediated by the electromagnetic *field* where electromagnetic *waves*, propagating with the speed of light, play a central role. This already brings to the fore two facts of fundamental importance in the description and explanation of natural phenomena - the role of the field as a basic dynamical entity, and the role of the speed of light as a physical quantity of overriding relevance.

When these two basic facts are considered together, the non-relativistic, or Newtonian view in physics proves to be deficient in providing for a consistent theoretical framework for the whole of physics. Instead, the *relativistic* point of view proves to be a more adequate one, where the space and time co-ordinates of an event (such as, a point particle being at a definite point in space at some definite instant of time) are all of similar relevance in describing the motion of a particle or a system of particles, and

time is transformed on a similar (though not quite the same) footing as the spatial coordinates in a transformation from one frame of reference to another.

However, the idea underlying the distinction between inertial and non-inertial frames remains pretty much the same: an inertial frame is one in which a particle free from all real forces, *including the gravitational* ones, moves uniformly. This scheme of things was given a new twist by Einstein who launched the idea that an inertial frame was one in which a particle free of all *non-gravitational* forces moves with a uniform velocity. This point of view has the consequence that as you move from one point in space to another where the gravitational field intensity is different, your inertial frame gets changed, i.e., in other words, the concept of the inertial frame is a *locally defined* one. What is more, moving over from one space-time point to another means a change in the way space-time events are related to one another in your frame. Put differently, the gravitational field at a point determines the local space-time *metric*, or the nature of the *terrain* you find yourself in, much like the way you move about being dependent on whether you are on the top a hill or in a deep ravine.

Thus, Einstein gave Newton's law of gravitation a completely new look whereby gravitation was conceived as describing the *geometry* of space and time, and what is perceived as a motion under a gravitational force in the Newtonian view, now appears as a 'free' motion in a changing landscape. Needless to say, this was more than just a new point of view, or a new way of describing things, since this new point of view made certain concrete predictions that were verified with reasonable certainty by specially designed experiments. Incidentally, the tying up of the concept of an inertial frame with the local gravitational field, implied a generalization of the Galilean principle of equivalence which now came to be referred to as the *general* principle of equivalence.

Einstein set up a complete set of equations describing the mutual effect of material particles and the gravitational field that permeates the whole of space and time.

This was the birth of the *general* theory of relativity which was a step forward compared to the *special* theory earlier developed by Einstein (see chapter 17 for a brief exposition

of the two theories), special in the sense of being a restricted one in that it did not include gravitation in its theoretical framework. However, even this general theory of relativity was a classical theory and hence was not a complete one since it did not incorporate quantum principles. Meanwhile, new ideas were being put together in another area where the theory of fields was being built up on the basis of quantum principles. In particular, the theory of the electromagnetic field in interaction with the fields corresponding to charged particles was put together where the point of view of the special theory of relativity was a necessary ingredient since the interactions were propagated at the speed of light and, at speeds nearing that of light, the Newtonian framework breaks down.

The quantum theory of fields (see sec. 16.15 for a brief introduction) that resulted from these considerations effected a great unification in the theoretical framework of physics, and culminated in the development of the *standard model* of elementary particle interactions. But problems continued to persist even in this remarkable theoretical framework.

For one thing, the standard model gave a unified explanation of the electromagnetic and the *weak* interactions, two of the four fundamental interactions of nature, while the gravitational and the *strong* interactions remained outside the purview of the theory. In particular, attempts to build up a quantum theory of gravity encountered deep problems. Quantum principles lead one to expect that the space-time structure predicted by the equations of the general theory of relativity should show up quantum *fluctuations* due to the fluctuations of the number of *gravitons*, particles postulated to represent energy quanta of the gravitational field, analogous to the photons, the latter being the energy quanta of the electromagnetic field. However, a theoretical scheme describing quantum fluctuations of the gravitational field and its interactions throws up inconsistencies at what is referred to as the *Planck scale*. The latter represents a set of inter-related critical thresholds of energy, mass, length, and time across which the general theory of relativity and the quantum theory of fields appear to be mutually incompatible.

A consistent quantum theory of gravitation is believed to be a necessary ingredient in a

unified explanation of the strong interaction forces along with the electromagnetic and weak ones. In other words, the development of such a theory could be a great event providing a unified theoretical framework for the whole of physics. A number of such unifying theoretical structures have been proposed awaiting their ultimate test in the theoretical and experimental domains.

Or, things may even develop along a totally unexpected and seemingly crazy course yet, where old frameworks dissolve and completely new ideas emerge. The older concepts retain a measure of validity in certain domains of experience while the emergent ideas set up a horizon as yet not dreamed of. Another thought, unsettling as it is, also deserves consideration: a *unified* explanation for the *whole* of physics may *itself* be an idealized concept bound up with the way we would *like* nature to confront us and which, at the end of the day, may prove to be an unfounded expectation after all. However, this is where physics has to give way to metaphysics and philosophy.

Chapter 6

Elasticity

6.1 Introduction: External and internal forces in a body

Think of any of the objects that you find to be in mechanical equilibrium around you. The forces exerted on it by various other bodies are termed *external* forces. For instance, the external forces acting on a chair resting on the floor are, first, the earth's gravitational pull on each and every particle making up the chair, the resultant of all these pulls being a single force acting through its center of gravity and, secondly, the reaction forces on its four legs exerted by the floor. The system of forces made up of these external forces is one in static equilibrium. If the body under consideration is imagined to be divided into a large number of small parts, then these external forces do not act uniformly on all the parts. For instance, in the above instance of the chair, the gravitational force acts in equal proportions on the various parts, but the reaction force acts only through the surfaces of contact of the legs with the floor.

Imagining now one of the small parts of the body, it is evident that the external forces acting on *this* part do not, in general, cancel one another, i.e., in other words, they do not form a system in equilibrium. How, then, can this part be in static equilibrium?

In this context, one needs to think of the *internal* forces arising due to the interactions among the small parts of the body under consideration. Every small part of the body

experiences internal forces exerted on it by other parts surrounding it, as well as some of the external forces due to the influence of other bodies. All these internal *and* external forces on the small part under consideration have to cancel one another in order that the part under consideration can rest in equilibrium. A further condition for equilibrium to be possible is that the *moments* of all these forces about any given point have to cancel one another.

The internal forces are electromagnetic in origin, caused by the interaction among the molecular and atomic charges in the material, and depend on the mutual positions of the small parts under consideration (many of the external forces are also of electromagnetic origin). With changes in these positions, the forces also get changed.

Now imagine that the external forces are made to change in some way. For instance, a weight may be placed on the chair in the above example. The small parts of the body cannot then continue to be in equilibrium in their previous positions. Since the external forces have changed, the internal forces must also get changed in order that all the small parts of the body can once again be in equilibrium. For some of the constituent parts of the body, the required changes in the internal forces may be comparatively large while for some other parts, the changes may be smaller. In any case, the mutual positions of the constituent parts have to undergo a change so as to bring about these changes in the internal forces.

6.2 Strain and stress

Thus, with a change in the external forces acting on the body under consideration, there occurs a change in the mutual positions of the constituent parts of the body, and an accompanying change in the internal forces among these parts. A new equilibrium configuration is achieved as the changes in the positions of the constituent parts and the consequent changes in the internal forces get adjusted in such a way that all the external and internal forces taken together, as also all their moments about any given point, cancel one another. The changes in the mutual positions of the small parts of the body under consideration is referred to as *strain*, while the consequent change in

the system of internal forces is termed *stress*. In the next section, I will give you the quantitative definitions of strain and stress at any given point in the body.

In general, one can say that the stress generated by the state of strain, i.e., the change in the system of internal forces, is of a *restitutive* and reversible nature. Thus, if the changes in the external forces causing the stress be reversed so as to restore the system of external forces to its configuration prior to the change, then the constituent parts also tend to get back to their previous positions, and the system of internal forces also tends to be restored. In the end, the previous configuration of equilibrium is restored. If, however, the magnitude of strain crosses a certain limit, a complete restoration of the earlier equilibrium configuration can no longer occur.

In mechanics, a rigid body is defined to be one for which the distances between all the particles making up the body are pairwise fixed and unalterable, which means that no strain can develop in such a body. This, however, is an idealization since in practice, all bodies are *deformable*. A deformable body can be in a state of strain, due to which stress forces are generated in it, the magnitude of the stress forces for a given state of strain depending on the material the body is made of.

We now come to the question of quantitative definition of strain and stress at any given point in a deformable body. Let us first look at the definition of strain.

Strictly speaking, the origin of strain and stress outlined above applies only for solid bodies since there is an important difference in this regard between a solid body and a liquid or a gas. In contrast to a solid in which the constituent particles (molecules) are, on the average, fixed in their respective positions, those in a liquid or a gas (a fluid in brief) are in incessant motion throughout the volume of the fluid. If we consider any two contiguous small volume elements in the fluid, a force is exerted on any one of these by the other, not only by virtue of the interactions of electromagnetic origin between the molecules of these two parts, but also by *momentum transfer* between them since molecules from one part enter into the other, carrying momentum with them. This transfer of momentum between contiguous parts of the fluid is responsible to a large extent for the internal forces occurring between various small parts of the

body.

For a solid body, one can imagine a configuration where there is *no* external force acting on it (in a weightless condition within a freely revolving artificial satellite, for instance), and consider the changes in the internal forces and relative positions of the small parts of the body with reference to this configuration. Evidently, the strain and stress will be zero for this reference configuration corresponding to zero external forces. For a fluid, on the other hand, such a configuration with zero external force is not conceivable since the fluid, in general, requires a containing vessel to keep it in equilibrium and there is always an external force acting on the fluid at the boundary where it is in contact with the vessel. Consequently, the fluid is always in a state of stress, since momentum transfer between the various parts of the fluids never ceases. Indeed, the stress in a fluid will be seen to be related to the *pressure* in it and, in the case of a fluid in motion, on the viscous forces of internal resistance. Since there exists no reference state for the fluid for which the strain and stress are zero in the same manner as in a solid, only the changes in the strain between given configurations are meaningful for it.

6.3 Quantitative definition of strain: strain parameters

Think of any given point, say, O in a deformable body and imagine a Cartesian co-ordinate system with O as the origin, whose co-ordinate axes are, say, OX , OY , OZ . Imagine a portion of the body in the shape of a small cube around O , whose edges are parallel to the co-ordinate axes, each edge being of length, say, a . Suppose now that a state of strain and stress has developed in the body in the vicinity of O . If the body be imagined to be made up of a large number of particles then the distances between these particles will be altered due to the strain, as a result of which the cube will undergo an alteration in shape and size. At the same time, the cube as a whole may suffer a translation and a rotation. But this translation and rotation of the cube as a whole will not concern us in defining the measure of strain since in none of these there occurs a change in the relative positions of the particles making up the cube. Such translations

and rotation of a body or a part thereof are referred to as *rigid displacements*.

Assuming that the alteration in the mutual positions of the particles making up the cube is small, the cube will, in general, be deformed to a parallelepiped due to the state of strain in the vicinity of O, and the lengths of the three adjacent sides of the parallelepiped will differ from the length (a) of the edges of the cube prior to the setting up of the strain. Under the above assumption of the alterations of mutual distances of the particles of the body near O being sufficiently small, the changes in lengths of the three adjacent sides will be proportional to a , and can be written as, respectively, ae_1 , ae_2 and ae_3 , where e_1 , e_2 and e_3 are three proportionality constants depending on the state of strain. Since these three quantities indicate the proportional changes in distances along the three axes, these give us a quantitative measure of the state of strain in the immediate vicinity of O.

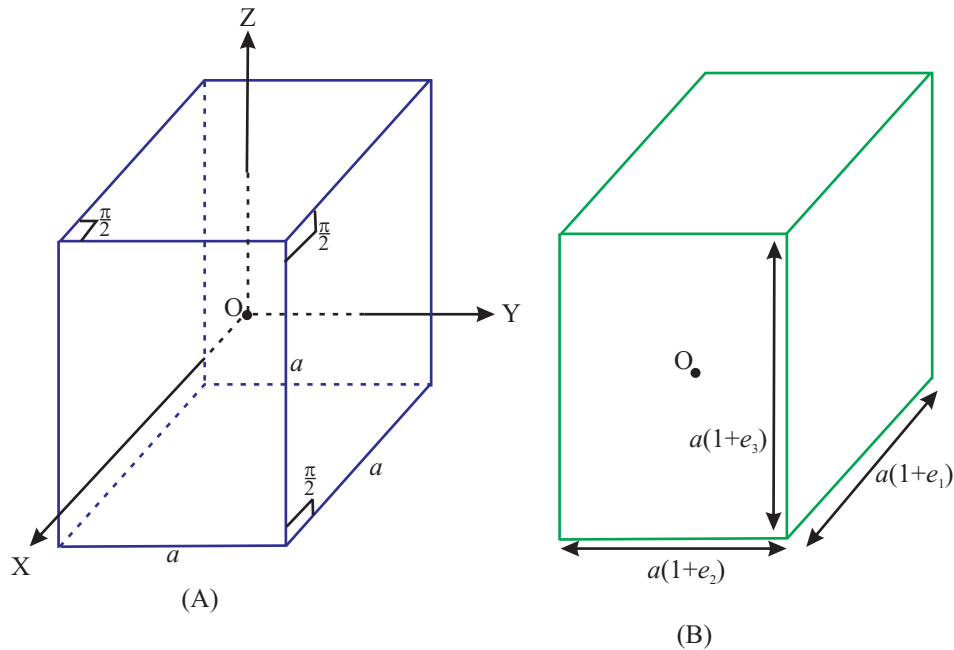


Figure 6-1: A cubical element of a deformable body around any given point O; the cube in (A) is transformed into a parallelepiped in (B) due to a small strain developed in the body around O; e_1 , e_2 , e_3 are the proportional changes in the lengths along the x-, y-, and z-axes respectively; these constitute three of the six strain parameters at O; the remaining three strain parameters (fig. 6-2) are not indicated.

However, not only the lengths of the sides of the cube, but also the *angles* between adjacent sides will, in general, be altered as result of the strain. Figure 6-1(A), (B) depict the cube and its altered form of a parallelepiped, while fig. 6-2(A), (B), (C) depict in a comparative manner, the sections, by the x-y, y-z, and z-x planes respectively, of the cube and of the parallelepiped.

Prior to the setting up of the strain, each of the angles between the three adjacent sides of the cube, considered pairwise, was $\frac{\pi}{2}$ while, in the strained configuration in the shape of the parallelepiped the angles are altered by, say, γ_1 , γ_2 , and γ_3 . For instance, the section of the parallelepiped by the y-z plane, will be a parallelogram shown schematically in fig. 6-2(A), with adjacent angles $\frac{\pi}{2} - \gamma_1$ and $\frac{\pi}{2} + \gamma_1$.

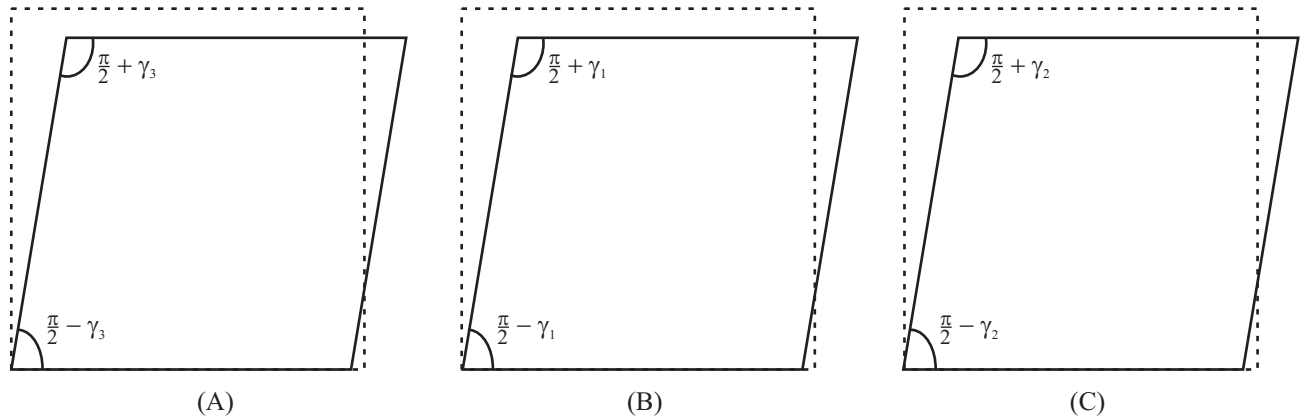


Figure 6-2: Sections of a cube, deformed into a parallelepiped, by (A) the x-y plane, (B) the y-z plane, and (C) the z-x plane; sections of both the cube (dotted) and the parallelepiped are shown for the sake of comparison where, in each case, one side of the square is shown to be coincident with the corresponding side of the parallelogram for the sake of comparison; each section of the parallelepiped is a parallelogram, and the strain parameters γ_1 , γ_2 , γ_3 are related to the angles of the parallelograms; the parameters e_1 , e_2 , e_3 of fig. 6-1 are not indicated.

The six quantities e_1 , e_2 , e_3 , γ_1 , γ_2 , γ_3 , which we collectively refer to as the strain parameters at O, describe completely the state of strain in the body in a small region around O. Evidently, a complete description of the state of strain at a point is not a very simple matter. In some simple situations, however, some of the six strain parameters may be zero, in which case the description of the state of strain may become simpler. We now turn to a few instances of this kind.

6.3.1 Tensile strain

Suppose that all the strain parameters excepting e_1 are zero. This means that in a region close to O, only the lengths along the x-axis suffer a change, and no other change in shape takes place. One then says that a *tensile strain* along the x-axis has developed in the body under consideration at the point (or close to) O. A cube imagined around O, with its edges along the three co-ordinate axes is transformed in this case into a rectangular parallelepiped, with its edges measuring, respectively, $a(1 + e_1)$, a , a . The proportional increase in length along the x-axis is here e_1 , which gives the quantitative measure of tensile strain. If it were the length along the y- or the z-axis that suffered a change instead of the length along the x-axis, then also the strain would have been of the same type where, in general, one has

$$\text{tensile strain} = \text{proportional increase in length} = \frac{\text{increase in length}}{\text{original length}}. \quad (6-1)$$

If any one edge of the cube were to get *decreased* in length, then its elongation would be *negative*, and so would be the tensile strain.

Homogeneous tensile strain

Till now I have talked of the state of strain in general, and tensile strain in particular, at any given point in a deformable body. It is generally the case that the strain parameters vary from point to point within the body. However, in special situations, the strain parameters may be the same throughout the body, in which case one says that a *homogeneous* or *uniform* strain has been developed in it. For instance, imagine a weightless deformable wire (an idealization; think of a wire of vanishingly small mass) to be fixed at the upper end, with a weight attached at the lower end, as in fig. 6-3.

In this case, the wire will be extended due to the load along its length in the vertical direction, and a tensile strain will develop in this direction. There will also occur a tensile strain in the horizontal direction at the same time, but we will not be concerned

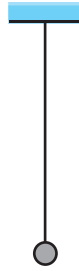


Figure 6-3: Weight suspended with a weightless deformable wire; the proportional increase of every small length element of the wire is the same regardless of its location in the wire, i.e., a homogeneous tensile strain is developed in it.

with of that for the time being. Since the wire has been assumed to be weightless, the tensile strain will be the same everywhere along the length of the wire, i.e., the proportional increase in length of any small length element in the wire will be the same regardless of its location in it (for a heavy wire or column, on the other hand, the strain will differ at different points along its length). Consequently, the tensile strain will be the same as the proportional increase in length of the *entire* wire, i.e., in this particular instance,

$$\text{tensile strain} = \frac{\text{increase in length of the wire}}{\text{initial length of the wire}}, \quad (6-2)$$

where the term initial length means length in the absence of the load. As mentioned above, for a wire whose mass is not negligible, the strain will no longer be homogeneous and then the tensile strain at any given point in it will not be the same as the proportional increase of the length of the wire as a whole.

6.3.2 Bulk strain

Consider a small volume element of a deformable body in the shape of a rectangular parallelepiped with edges parallel to the three axes of a chosen rectangular co-ordinate system, the lengths of the edges being a_1, a_2, a_3 , and assume that it gets transformed, in a deformation of the body, to a rectangular parallelepiped of slightly altered edge lengths, with the edges still parallel to the co-ordinate axes. This means that there occurs no change in shape of the element involving a change in the angles between

adjacent sides. One then says that a *bulk strain* (or *volume strain*) (along with tensile strains) has developed in a small region in (or around) the point (O, fig. 6-1) under consideration, where the bulk strain is attended, in general, with tensile strains along the three co-ordinate axes (thus, the situation considered in sec. 6.3.1 is a special case, where two of the tensile strains are zero). The quantitative measure of bulk strain is here given by

$$\text{bulk strain} = \frac{\text{increase in volume}}{\text{initial volume}}. \quad (6-3)$$

Assuming that the edge lengths of the parallelepiped become $a_1(1+e_1)$, $a_2(1+e_2)$, $a_3(1+e_3)$ due to the strain, where e_1, e_2, e_3 denote the proportional increases in the lengths, one finds that the increase in volume is

$$a_1 a_2 a_3 (1+e_1)(1+e_2)(1+e_3) - a_1 a_2 a_3 \approx a_1 a_2 a_3 (e_1 + e_2 + e_3), \quad (6-4)$$

where the proportional changes in length (e_1, e_2, e_3) , i.e., the magnitudes of the tensile strains, are assumed to be small, in which case one has

$$\text{bulk strain} = e_1 + e_2 + e_3 = \text{sum of tensile strains along the three directions}. \quad (6-5)$$

In other words, a bulk strain can be expressed in terms of three independent tensile strains. The bulk strain is negative if there occurs a contraction of volume instead of an expansion.

Strictly speaking, all our considerations apply only in the limit of vanishingly small strains. It is only for such small strains that the proportionality between strain and stress, to be introduced later, holds. One then needs to retain in calculations relating to strain and stress, only the first degree terms in the strains as in eq. (6-4).

6.3.3 Shear strain

Suppose that a small cube imagined around a point O in a deformable body undergoes a change of shape, but no change in the edge length, due to the strain developed in

it and that the section by the y-z plane gets transformed from a square to a rhombus while, more generally, the section assumes the shape of a parallelogram. This is similar to fig. 6-2(B) where both the sides of the parallelogram are to be assumed to be of the same length, namely, a since the strain has been assumed to be such that there is no change in the linear dimensions of the cube, the only change being one of shape. Thus, one has $e_1 = e_2 = e_3 = 0$ in this case, implying that there is no tensile strain along any of the co-ordinate axes, and also no bulk strain, at the point O.

The strain in this case is, moreover, such that only the angles between the adjacent sides in the y-z section have suffered a change, there being no change in the angles in the x-y and the z-x sections. Looking at fig. 6-2(B), the angles are, say, $\frac{\pi}{2} - \gamma_1$ and $\frac{\pi}{2} + \gamma_1$.

The strain in this case is said to be in the nature of a *shear* in the y-z plane, the quantitative measure of the shearing strain being

$$\text{shear in the } yz \text{ plane} = \text{decrease in angle between the edges along the } y \text{ and } z \text{ axes} = \gamma_1. \quad (6-6)$$

In this case one has, moreover, $\gamma_2 = \gamma_3 = 0$, i.e., only one of the six strain parameters, namely, γ_1 , is non-zero. Had the shear taken place in the z-x or the x-y plane instead of the y-z plane, one would have had a similar situation, and the shearing strain would also be defined in a similar manner.

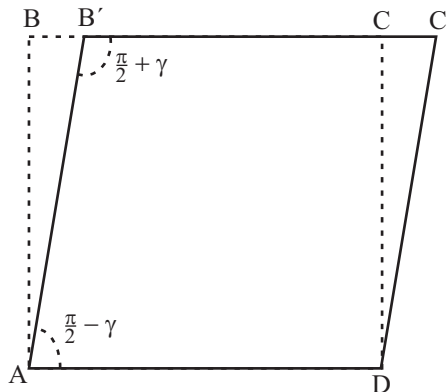


Figure 6-4: Illustrating the alternative definition of shearing strain as in eq. (6-7).

Fig. 6-4 depicts a shearing strain in the plane of the drawing, which is similar to 6-2(A), (B), or (C). The difference between the situations depicted in fig. 6-4 and any one of figures 6-2(A), (B), (C) resides in the fact that, in the former, the tensile strain is taken to be zero, as a result of which the section of the cube under consideration is deformed to a rhombus rather than a parallelogram. Since we assume that all strains are small, $B'C'$ may be taken to be coincident with the line along BC. Here the shearing strain (γ in the figure) can alternatively be defined as

$$\text{shearing strain} = \frac{\text{displacement of BC parallel to AD}}{\text{distance between AD and BC}}. \quad (6-7)$$

Analogous to tensile strain, the shear strain can also be either positive or negative. Consider, for instance, the shear in the y-z plane where the y-z section of the cube gets deformed into a parallelogram, as in fig. 6-2(B). Referring to the edges parallel to the y-axis, if the edge with the larger value of the z- co-ordinate is displaced (relative to the other edge) along the direction of increasing y- co-ordinate, then the strain is defined to be positive, while the strain is negative if the displacement is in the opposite direction.

Problem 6-1

The square shaped cross-section of a block gets deformed due to a pure shear. If the edge length of the cross section is $a = 0.2$ m and the shear strain in the plane of the cross-section is $\gamma = 0.02$, find the lengths of the two diagonals of the cross-section of the deformed block.

Answer to Problem 6-1

HINT: As a result of pure shear, the cross section becomes a rhombus with angles $\frac{\pi}{2} - \gamma$ and $\frac{\pi}{2} + \gamma$. The length l_1 of the larger diagonal is given by $l_1^2 = 2a^2(1 + \sin \gamma)$ (check this out; use the geometry of the rhombus), i.e., $l_1 \approx \sqrt{2}a(1 + \frac{\gamma}{2}) = \sqrt{2} \times 0.2 \times 1.01$ m. The shorter diagonal is of length $l_2 \approx \sqrt{2} \times 0.2 \times 0.99$ m.

6.3.4 Mixed strain

We have thus met with a number of situations where the state of strain is described in simple terms. In general, however, the strain produced at a point in a deformable body is more often than not of a more complex nature where none of the six strain parameters ($e_1, e_2, e_3, \gamma_1, \gamma_2, \gamma_3$) is zero. Such a situation is described by saying that a *mixed* strain has developed at the point under consideration. A mixed strain involves, in general, tensile strains along all directions, shearing strains in all planes, and bulk strain. In quantitative terms, the tensile strains along the x-, y-, and z-axes are respectively e_1, e_2, e_3 , the shearing strains in the y-z, z-x, and x-y planes are, respectively, $\gamma_1, \gamma_2, \gamma_3$, and the bulk strain, calculated according to formula (6-3), is $e_1 + e_2 + e_3$.

6.3.5 Principal axes. Principal components of strain.

It is important to note that if the co-ordinate axes at the point under consideration were chosen differently, i.e., in other words, if the edges of the small cube were oriented along a different set of directions, then the *description of strain would have been different*. Instead of the strain parameters $e_1, e_2, e_3, \gamma_1, \gamma_2, \gamma_3$ one would then have had, a different set of parameters, say, $e'_1, e'_2, e'_3, \gamma'_1, \gamma'_2, \gamma'_3$. However, the two sets of parameters have to be related to each other in some definite manner since they represent the *same* state of strain at the given point.

In particular, one can choose a set of axes for which the description in terms of the new set of strain parameters is quite simple: in this new set, $\gamma'_1, \gamma'_2, \gamma'_3$ are all zero. In other words, any state of strain can be described as a combination of three tensile strains along three *specially* chosen Cartesian axes, and an associated volume strain. These are termed the *principal axes of strain* at that point, while the three tensile strains, which we denote as $e_1^{(0)}, e_2^{(0)}, e_3^{(0)}$, are termed the *principal components* of strain.

The transformation from one set of strain parameters to another due to a change of co-ordinate axes is reminiscent of the transformation of the components of a vector. Indeed, the six strain parameters collectively stand for a single physical quantity, just as the three components of a vector represent a single physical quantity like, for instance,

the velocity of a particle. The quantity represented by the six strain parameters is termed the strain *tensor*. Tensors can be classified into those of *rank* one, two, three, etc., where the rank determines the number of independent components required for a complete description of the quantity. Tensors of rank one are just the vectors we are familiar with. The strain tensor is a *symmetric* tensor of rank two, requiring six components for its complete description. Symmetric and *antisymmetric* tensors are of special types while, in general, a tensor is made up of a symmetric and an antisymmetric part. Finally, scalars can be looked upon as tensors of rank *zero*.

Problem 6-2

A homogeneous rod of length $l_0 = 0.5$ m with a square cross-section of edge length $a_0 = 0.03$ m suffers a deformation with tensile strain $\epsilon_1 = 0.01$ along its length and $\epsilon_2 = -0.002$ along each of the two edges of its cross-section. Estimate the changed volume of the rod.

Answer to Problem 6-2

HINT: The changed length is $l = l_0(1 + \epsilon_1)$ while the changed edge lengths are $a = a_0(1 + \epsilon_2)$ each. Thus the changed volume is $v = la^2 \approx l_0 a_0^2 (1 + \epsilon_1 + 2\epsilon_2) = 0.5 \times (0.03)^2 \times 1.006 \text{ m}^3$.

6.4 Stress in a deformable body

As I have pointed out, the mutual positions of the constituent particles of a deformable body get altered as a strain is developed in it. Due to these altered mutual positions, the internal forces arising due to the mutual interaction among these particles are also changed, and these changes are, in general, restitutive in nature. Another characteristic of the internal forces is that these are *short range* ones: as the distance between a pair of particles is made to increase, the interaction force between them decreases rapidly.

In other words, imagining a small element of volume within the body, only the particles located within a small neighborhood of that element can exert internal forces on it, while the forces exerted by particles located further away are negligibly small. Consequently,

if one imagines two small elements adjacent to each other then the internal force due to one element on the other, which acts through their common boundary, will be proportional in magnitude to the surface area of that boundary. A force of such a nature is termed a *surface* force.

1. By contrast, if the force on a volume element is proportional to the volume of that element then one describes it as a *volume* force. An example of such volume force is provided by the gravitational force, though the latter is, generally speaking, too weak to be of relevance in a consideration of the elastic properties of a body. Generally, from a mathematical point of view, a surface force can be represented in the form of a volume force, but a volume force cannot always be represented as a surface force. In other words, the internal forces mentioned above can be expressed either as surface or as volume forces, the former being the commonly adopted mode of description. Though this is an interesting aspect of the internal forces, I will not pursue it further. It is the system of internal forces that generates the state of stress at a point in a deformable body.
2. All these above considerations, are applicable for a deformable solid and, with a few qualifications, for a fluid as well. As I have indicated above, the internal force exerted by one part of a fluid on a contiguous part is largely due to the momentum transfer from the former to the latter by virtue of molecular transport. Since the transport occurs through the common surface of the two parts, the mutual force is once again proportional to the surface area, and is thus a surface force.

In fig. 6-5, 'abcd' represents a small part of a surface around any given point O, separating two regions, A and B, in a deformable body (the other boundary surfaces of the regions are not shown in the figure). The force exerted by the region B on A due to the interaction between the particles of these two regions, exerted through this part of the common surface, is a vector quantity and has been shown by the double-headed arrow in the figure. An equal and opposite force is exerted by A on B.

The direction and magnitude of this force, exerted by B on A through the surface element 'abcd', *per unit area* of the surface, depend on two things: (a) on the relative positions of particles in the vicinity of the point O which determine the interaction forces between

these particles, i.e., on the *state of strain* at O, and (b) on the orientation of the element 'abcd', i.e., in other words, on the direction of the normal drawn to this element - this has been shown with a single-headed arrow in the figure.

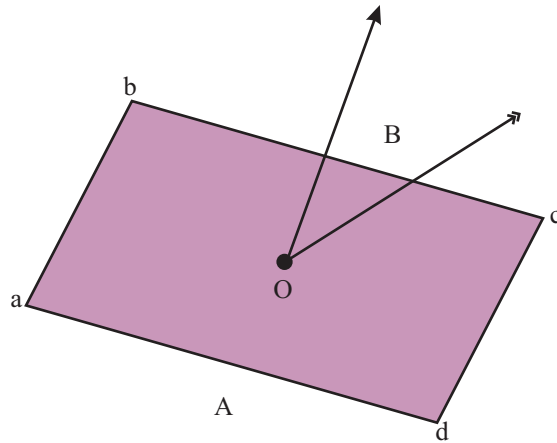


Figure 6-5: Illustrating the idea of stress at a point; regions A and B within a deformable body have a common surface of separation, a part of which ('abcd') around the point O is shown; the internal force exerted by B on A through this surface element is shown with a double-headed arrow, while the normal to the surface at O is shown with a single-headed arrow.

Referring to the special cases where the element 'abcd' is imagined to be parallel to the x-y, y-z and z-x planes respectively, and considering in each case the three components of the force per unit area, one obtains a total of nine quantities. However, not all of these are independent of one another, only six of the nine quantities being mutually independent.

This follows from the requirement that the stress forces are to be in the nature of surface forces.

Knowing these six quantities, one can then obtain from these the force for any *other* orientation of the surface element 'abcd'. In other words, these six quantities give a complete description of the system of internal forces in the vicinity of the point O. The internal forces between contiguous parts in a close vicinity of the point O are said to constitute the state of *stress* at O. Thus one can say that the complete quantitative

description of the state of stress at O is provided by these six quantities. This set of six quantities, providing us with a measure of stress may, in general, vary from one point to another in the body under consideration.

When I say ‘a small neighborhood around O’ or ‘in the close vicinity of O’, I actually mean a region of vanishingly small extent around O. The quantities measuring strain and stress at O give us a complete description of the changes in mutual positions of the particles and of the internal forces in such a vanishingly small region.

Evidently, the quantitative description of the state of stress at a point is not a very simple matter, this being similar to what we found in the quantitative description of strain. I will therefore address first a number of special situations where the stress can be defined in a simple way.

6.4.1 Tensile stress

Following the above paragraphs, think again of the regions A and B on the two sides of the surface element ‘abcd’ (fig. 6-5) within a deformable body. If the internal force exerted by B on A through this area element happens to be along the normal to it (we assume the normal to be pointing in the direction *from A to B*), then we will say that a tensile stress has developed at O in the direction of the normal. If this force is denoted by δF (this force being reckoned *negative* if it happens to be directed from B to A) and if the area under consideration is δS , then the quantitative definition of tensile stress is expressed as

$$\text{tensile stress} = \frac{\delta F}{\delta S}. \quad (6-8)$$

In brief, then, one defines tensile stress as the internal force acting per unit area in the direction of the normal to a surface around the point under consideration, in the sense of the outward normal relative to the element of the deformable body on which it acts. Imagining the normal to point along the x-, y- and z-axis respectively, we find that *three*

independent tensile stresses may develop at a point in a deformable body.

6.4.2 Shear stress

Suppose, in fig. 6-5, that the internal forces are such that the force (δF) exerted by B on A is directed parallel to the surface element 'abcd'. One then says that a *shear* stress (or shearing stress) has developed at O across the plane 'abcd'. The quantitative definition of shear stress is of the same form as (6-8):

$$\text{shearing stress} = \frac{\delta F}{\delta S}. \quad (6-9)$$

One has to remember, though, that the directions of δF with reference to the orientation of δS in equations (6-8) and (6-9) are different.

Imagining the element 'abcd' to be parallel to the x-y, y-z, and z-x planes respectively, one obtains three independent quantities of the above description, which means that there can be three independent shear stresses at any point O in a deformable body.

When the surface element is parallel to the x-y plane, for instance, the force δF may have two independent components, and it may then appear that, considering the three possible independent orientations, there will be in all six independent quantities describing shear stress at O. However, of these six, only three can be mutually independent if the net torque on any volume element around the point under consideration is to be zero.

This is related to the fact that the stress forces are in the nature of surface forces, which implies restrictions on the moments of the stress forces acting on a volume element, about a point chosen within that element.

Thus, in conclusion, one arrives at six independent quantities, of which three relate to tensile stress and three to shear stress. Making use of these six, one can give a complete description of the state of stress at any given point in a deformable body, regardless of

the whether the stress is of a simple kind or not. These are precisely the six quantities I mentioned while talking of stress at a point.

Analogous to the six strain parameters, the six quantities characterizing the state of stress at a point, constitute a symmetric tensor of rank two.

Problem 6-3

Forces of magnitude F each are applied on two opposite faces of a homogeneous cubical block of edge length l , as in fig. 6-6 in which a cross-section ABCD of the cube is shown. Find the tensile stress and the shear stress on the diagonal face BD.

Answer to Problem 6-3

HINT: Consider the upper triangular part (ABD in cross-section) of the cube. The internal force exerted on this part by the lower triangular part BCD must be F perpendicular to the face AB and acting through O, the mid-point of BD if the upper part is to be in equilibrium (reason out why; it is assumed that the forces applied on the faces AB, DC act through the mid-points of these faces; more generally, the forces are distributed over the surfaces). The resolved parts of this force along directions parallel and perpendicular to the diagonal face BD are of magnitude $\frac{F}{\sqrt{2}}$ each. Since the area of cross section of the diagonal face is $\sqrt{2}l^2$, the required tensile and shear stresses are $\frac{F}{2l^2}$ each.

6.4.3 Bulk stress

Suppose that the stress developed at a point O in a deformable body is such that the tensile stresses along the x-, y- and z-axes are equal and, moreover, the shear stresses in the x-y, y-z, and z-x planes are all zero. One then says that a bulk stress (or *volume stress*) has developed at that point, and the measure of bulk stress is taken to be any of the three tensile stresses at the point. More generally, if the three tensile stresses are unequal, one can describe the state of stress at O as a mixture of a bulk stress and three tensile stresses (refer to sec. 6.4.4). In this case, if s_x, s_y, s_z be the tensile

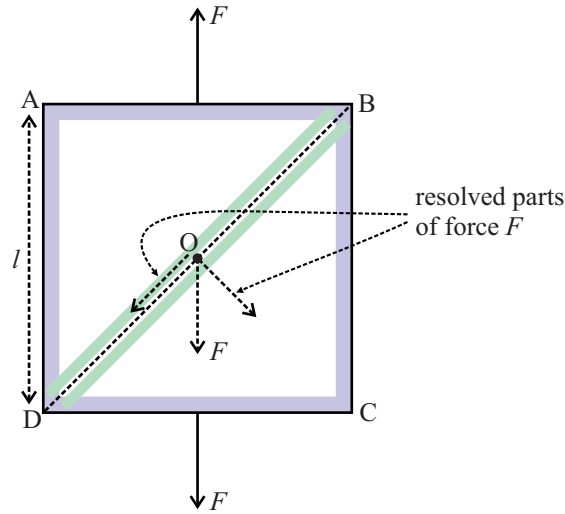


Figure 6-6: A homogeneous cubical block of edge length l shown in cross-section (ABCD); forces of magnitude F each are applied to the two faces AB, DC along a direction perpendicular to each face; the internal force exerted by the lower triangular part BCD on the upper part ABC is F (dotted arrow) acting along a direction perpendicular to AB through O, the mid-point of BD. This results in a tensile stress and a shear stress on the face BC, each of magnitude $\frac{F}{2l^2}$ (see problem 6-3).

stresses along the three co-ordinate axes (we assume for the sake of simplicity that there is no shear stress in any of the three co-ordinate planes), then the state of stress can be described as a volume stress $s = \frac{1}{3}(s_x + s_y + s_z)$ mixed with tensile stresses $s'_x = s_x - s$, $s'_y = s_y - s$, $s'_z = s_z - s$ along the three axes.

6.4.4 Mixed stress: principal components of stress

In a simple situation like, say, the one where only a tensile stress is developed along one of the three co-ordinate axes, only one of the six stress parameters is non-zero. In the case of a pure bulk stress there are three tensile stresses involved, but all the three stresses are equal. More generally, however, none of the six stress parameters is zero and all of them are unequal. This is the case of *mixed stress* at the point under consideration. However, analogous to the description of mixed strain, one can choose a special set of axes relative to which the stress is described completely as a combination of three tensile stresses. These are termed the *principal stress components* at the point and the special set of axes is referred to as the *principal axes of stress*.

6.5 Stress-strain curve

6.5.1 A weightless wire with a load

Consider the weightless wire of fig. 6-3. Let the initial length of the wire be L and let the elongation due to the applied load W be l . If the area of cross section of the wire (assumed homogeneous everywhere) be A , then the tensile stress at any point in it will be $S = \frac{W}{A}$, i.e., in other words, if the weight W is made to increase, then the tensile stress will also increase proportionately. Here we ignore the slight change in the area of cross-section of the wire due to the applied load.

In reality, however, a change in the cross-section does take place (see sec. 6.6.2) and has to be taken into account in certain considerations.

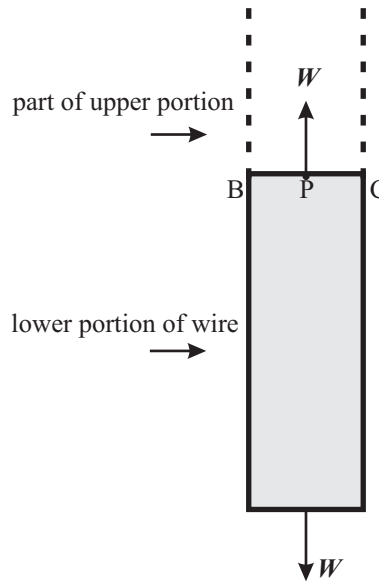


Figure 6-7: Explaining the calculation of stress at a point in a weightless wire with a load W applied at its lower end; a portion of the wire at the lower end is shown, while a portion above this is shown with dashed lines; BC is a cross-section of the wire at the point P, across which a stress force W is developed so as to keep the portion under consideration in equilibrium.

In order to arrive at the above result relating to the stress in a weightless wire due to an applied load, look at fig. 6-7. In this figure, the cross-section is magnified for the sake

of convenience. BC is the cross-section of the wire at any given point P in it, and the lower part of the wire from P downward is shown, while a part above this is shown with dashed lines. Considering the equilibrium of the lower part, one finds that the internal force exerted on this by the upper part has to be W in the vertically upward direction for this lower part to be in equilibrium, since the only other force acting on this part is the force due to the weight W acting vertically downward (recall that the wire itself has been assumed to be weightless). Thus, the stress developed at P is a tensile one, there being no shear stress involved since the internal force has no component parallel to the cross-section BC. The expression for the tensile stress, according to (6-8) is

$$s = \frac{W}{A}, \quad (6-10)$$

which shows that the stress is everywhere the same in the wire. Note that this would no longer be true for a heavy wire or rod.

While the tensile stress at any point in the wire is proportional to the load W , the tensile *strain* at any point is similarly proportional to the elongation l of the wire as a whole. The fact that the strain is uniform throughout the wire follows from the stress-strain relations to be introduced later. By definition, the tensile strain (ϵ) is equal to the proportional increase in length of the wire as a whole (reason this out, assuming that the strain is uniform, see eq. (6-2)):

$$\epsilon = \frac{l}{L}. \quad (6-11)$$

6.5.2 The curve: principal features

Performing an experiment with different values of W and measuring the corresponding values of the elongation l , one may work out the stress and strain values and then plot a *stress-strain curve*. It turns out to be worthwhile to distinguish between the *nominal* and *true* values of the stress and strain, where the former are calculated with the cross-section and length of the wire one starts with, and the latter with the actual cross-section and length at any given stage of the experiment.

Analogous to the graph showing the relation between the tensile stress and tensile strain, one can have graphical plots for other components of the stress and strain tensors as well. All such stress-strain curves (where true stress is plotted against true strain), drawn for various materials, have a number of qualitative features in common. However, they differ greatly in details. Even for two bodies made of the same material, the curves may differ, depending on the past history of the bodies. For our present purpose, I will concentrate on the common features observed in the various stress-strain curves, referring to fig. 6-8, which gives a schematic representation of the stress-strain relation.

A stress-strain curve is, in the main, made up of two parts, a part corresponding to *elastic deformation*, and one corresponding to *plastic deformation*. In the figure, the former corresponds to the part from O to A while the latter to the one from A to B. However, the transition from elastic to plastic deformation cannot be defined uniquely and so, the location of the point A, referred to as the *elastic limit* is not very precise. What distinguishes the elastic from the plastic deformation is that the former is *reversible* while the latter is *not*. In other words, referring to the loaded wire for instance, when the load is removed the wire returns to its former length and cross-section if the deformation happens to be an elastic one. More generally, if the material under consideration is deformed by the application of external forces and then these external forces are removed so as to cause the stress to return to zero value, the strain becomes zero at the same time, provided the deformation is kept within the elastic limit. Beyond the elastic limit, on the other hand, decreasing the stress to zero value causes a *residual strain* to remain in the body.

This is shown schematically by the dotted line CD in fig. 6-8 which tells us that as the body is deformed from O to C (beyond the elastic limit A) and then the stress is made to decrease to zero, the variation of strain occurs not in the reverse sequence from C to O through A, but along CD, resulting in the residual strain corresponding to OD.

The relation between stress and strain is found to be a relatively simple one from O up to the elastic limit A, being more or less *linear* though, in reality, the graph may be

slightly bent as one approaches the elastic limit. Beyond A, however, the relation is of a more complex nature and a unified description of the relation turns out to be difficult.

1. The distinction between elastic and plastic deformations holds for solid bodies. For a fluid, plastic deformations are not of relevance.
2. Strictly speaking, plastic deformation starts from the very beginning of the loading sequence. However, for most practical purposes, it remains negligible up to the elastic limit. The distinction between elastic and plastic deformations is more a matter of practical description than one of principle.

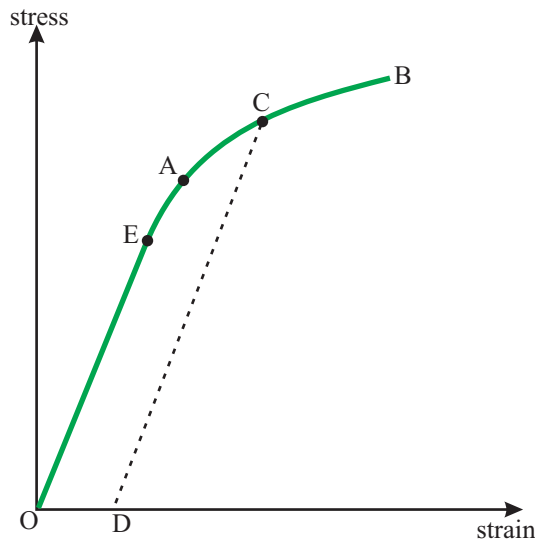


Figure 6-8: Stress-strain curve (schematic); the curve consists of two parts - the part from O to A, which is more or less linear, corresponds to elastic deformation while that from A to B depicts plastic deformation; the point A corresponds to the elastic limit; the plastic deformation is irreversible in that, removing the stress leaves a residual strain in the material (represented by OD if the stress is removed on reaching the point C in the deformation process); fracture of the material occurs at B; within the elastic limit, proportionality between stress and strain holds up to E, while the curve gets slightly bent beyond E.

Problem 6-4

A cable of cross-sectional area $3.5 \times 10^{-4} \text{ m}^2$ can bear a stress of $2.4 \times 10^8 \text{ N} \cdot \text{m}^{-2}$ before the elastic limit is reached. What is the maximum permissible upward acceleration of a 2000kg elevator pulled up by the cable if the stress in the latter is not to exceed one third of the elastic limit?

$$(g = 9.8 \text{ m} \cdot \text{s}^{-2}.)$$

Answer to Problem 6-4

HINT: If T denotes the tensile stress force in the cable, then the net upward force on the elevator is $T - mg$ where m ($= 2000 \text{ kg}$) stands for the mass of the elevator. This provides for the acceleration (f) of the elevator (referred to a frame fixed to the ground). Using the maximum permissible value of T ($= \frac{1}{3} \times 2.4 \times 3.5 \times 10^4 \text{ N}$), one obtains the required maximum allowed acceleration as $f_{\max} = \frac{T_{\max} - mg}{m} = 4.2 \text{ m} \cdot \text{s}^{-2}$.

Problem 6-5

A block of mass $m = 10 \text{ kg}$ rests on an inclined plane making an angle $\theta = \frac{\pi}{3}$ with the horizontal. The block is supported, so as to just prevent sliding, by a wire of cross-sectional area $\alpha = 1.0 \times 10^{-6} \text{ m}^2$ pulled with a force T , where the wire makes an angle ϕ with the inclined plane (see fig. 6-9); the coefficient of friction between the block and the inclined plane is $\mu = \frac{1}{\sqrt{3}}$; it is found that the wire snaps when the angle of pull ϕ is increased from a low value to $\phi_{\max} = \frac{\pi}{6}$. Find the breaking stress for the wire under stretching ($g = 9.8 \text{ m} \cdot \text{s}^{-2}$).

Answer to Problem 6-5

HINT: the forces acting on the block are its weight mg acting vertically downward, the force of pull T along the wire, the normal reaction N (say) exerted by the inclined plane, and the force of friction μN preventing sliding (see fig. 6-9). The conditions of equilibrium, as obtained by resolving the forces along the plane and perpendicular to it are, respectively, $T \cos \phi + \mu N = mg \sin \theta$, and $T \sin \phi + N = mg \cos \theta$; from which one obtains $T = mg \frac{\sin(\theta - \theta_0)}{\cos(\phi + \theta_0)}$, where $\theta_0 = \arctan \mu$ (shown in the figure) is referred to as the angle of friction, being the maximum inclination of the plane to the horizontal for which no support is necessary. The maximum value of ϕ for which the wire snaps is then given by $\cos(\phi_{\max} + \theta_0) = \frac{mg}{S_0 \alpha} \sin(\theta - \theta_0)$, where the breaking stress $S_0 = \frac{T_{\max}}{\alpha}$ has been made use of, T_{\max} being the maximum possible force of pull before the wire snaps. Substituting given values, one obtains $S_0 = 9.8 \times 10^7 \text{ N} \cdot \text{m}^{-2}$.

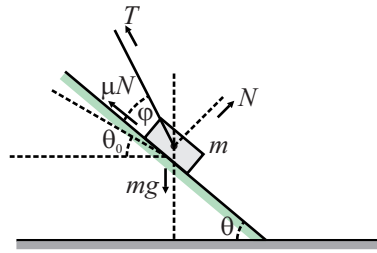


Figure 6-9: A block of mass m resting on a plane inclined to the horizontal at an angle θ ; the coefficient friction between the block and the plane is μ , where $\tan \theta > \mu$; the block is supported, so as to just prevent sliding, by a wire of cross sectional area α , with a force T , where the direction of pull makes an angle ϕ with the inclined plane; the wire snaps as the angle ϕ is made to increase from a low value up to a certain value ϕ_{\max} , depending on the breaking stress S_0 in the wire - refer to problem 6-5; no supporting force by means of the wire is necessary if $\tan \theta < \mu$; the angle $\theta_0 = \arctan \mu$, referred to as the angle of friction, is shown.

6.5.3 Elastic and plastic deformations

While I have distinguished between the elastic and plastic deformations by referring to the *experimentally* observed behavior of a deformable body, it would be far more desirable to distinguish between the two in terms of the *deformation processes occurring in the material* of the body. For this, I refer to a *crystalline solid* such as a metal, since crystalline solids have been studied thoroughly in the context of stress-strain relations.

The atoms in a crystalline solid are arranged in a regular pattern, forming a lattice. In the lattice, the atoms are held together by mutual forces, much like a set of balls connected by springs. While the atoms are in a state of continual thermal vibration, one can, for the present purpose, ignore this thermal motion and assume that the atoms are held fixed in their respective mean positions. As the material is strained to a small extent, the atoms get shifted from their previous positions, with a consequent change in the mutual forces between them, similar to changes in the force due to springs as they are extended or compressed. Here the springs represent the bonds between the atoms, and a small strain corresponds to the springs getting extended or compressed. Analogous to the extension or compression of a spring being reversible and the spring force being proportional to the extension or the compression, the deformation is also reversible for such small strains, and the stress-strain relation is predominantly a linear one.

For relatively larger deformations in a spring, the spring force becomes *non-linear* or *anharmonic* in nature while the process of deformation still remains reversible. This explains the bending of the stress-strain curve away from a linear shape for relatively larger deformations within the elastic limit (E to A in fig. 6-8).

This spring model constitutes a more or less adequate explanation of the nature of deformation up to the elastic limit of a material. Within this limit, the changes in bond lengths and orientations are the dominant processes in terms of which the stress-strain relation can be explained. In reality, there does occur a *snapping* or breaking of some of the inter-atomic bonds as well, but such events are rare when the deformation is small, and do not show up in the *macroscopically* determined stress-strain relations. Finely tuned experiments, however, can detect such snapping of bonds in the material. Along with bonds being snapped, *new* bonds are established between contiguous atoms in the crystal as the deformation is continued. These events of breaking and making of bonds constitute, to a large extent, an *irreversible* process whereby some energy is lost to vibrational modes of neighboring atoms, and the bonds cannot be re-established in exactly the reverse sequence compared to the one followed during the straining process. One expresses this by saying that bond-breaking is associated with an *absorption* of energy. Thus, a small degree of irreversibility and plasticity remains in the deformation of a crystalline sample even within the elastic limit, but this can be disregarded for most practical purposes.

The term 'snapping' need not be taken in the literal sense of snapping of a thread or a rope. While a bond stands for a configuration of atoms or molecules characterized by a low value of the energy, a configuration of substantially higher energy may be described as one where the bond has been 'snapped'. A special case corresponds to the transition from a bound to an unbound configuration, which is more like the snapping of a thread. In the process of plastic deformation, bonds are not completely broken, but are loosened to an appreciable extent, and are then replaced with relatively tight ones.

Beyond the elastic limit, it is this breaking (and making) of bonds that assumes a greater degree of importance, and shows up in macroscopic measurements of stress and strain. Theoretical estimates show that, in a perfectly formed crystal, comparatively large stress forces are required before bond-breaking can occur on a large scale because this requires a large number of bonds to be broken *simultaneously*. This, however, is far from the case for a *real* crystal where *imperfections* in the regular arrangement of atoms are always there. While there can be various different *types* of imperfection, one commonly occurring type is a *dislocation* where a crystalline plane made up of regularly arranged atoms gets terminated along a line, with the other half of the plane missing (fig. 6-10). Looking at any point on this line, the bonds between neighboring atoms near this point get stretched and compressed (not shown in fig. 6-10; only the stretching on one side of the dislocation line can be inferred from the figure) so as to accommodate the faulty arrangement of atoms. And it is relatively easy to create a deformation of the crystal by a step-by-step movement of the dislocation (say, from left to right in the figure) where, at each step, only a relatively small number of bonds are broken and formed.

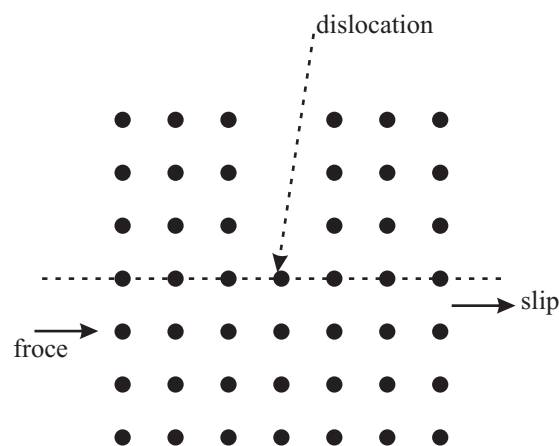


Figure 6-10: Illustrating a dislocation in the regular arrangement of atoms in a crystal; a row of atoms (belonging to a plane perpendicular to the plane of the figure) gets terminated at a point, the atoms on the other side of the point being missing; on the application of a shearing force, there occurs a slip between planes above and below the dislocation and the dislocation propagates by the breaking and remaking of bonds between neighboring atoms around it; the applied force causing the shear operates on atoms both above and below the slip plane.

What is of importance here is the fact that the strain and stress in the crystal can be

produced in relatively small steps due to the motion of the dislocations while, at the same time, the process is irreversible since it involves, along with the stretching and compression of the bonds, the snapping of bonds as well. The motion of a dislocation corresponds to a slipping of one atomic plane over another and is essentially of the nature of a shear within the crystal, wherein the volume of the crystal remains, to a good degree of approximation, unaltered. For any given material, it requires a certain minimum stress to be developed before this slipping of one atomic plane over another is initiated, which determines the elastic limit of the material. This limit depends on the density of the pre-existing dislocations in the crystalline material - since it is the movement of the dislocations that causes the irreversible deformation, the elastic limit is large for a material with only a few dislocations in it.

A puzzling question that now comes up is the following. Once the dislocations start moving, what prevents their continued motion under a given stress, provided the requisite amount of energy is supplied to the material? The answer to this lies in the observation that the motion of a dislocation leads to further dislocations being created in the crystal owing to the dislocation line meeting a densely populated crystal plane in which it is easy for a dislocation to be produced. When the density of dislocations in a certain region in the crystalline material becomes relatively high, a self-limiting process of one dislocation impeding the motion of another sets in. As a result, the straining process stops, and a further increase in the stress becomes necessary to generate more deformation. This is the process commonly referred to as *cold-hardening* or *work-hardening* of a material.

In summary, then, the stress-strain curve of a material essentially consists of two parts, one corresponding to elastic deformation and the other to plastic deformation. While the elastic deformation corresponds to an essentially linear stress-strain relation and is reversible in nature, the plastic deformation is irreversible, with a residual strain remaining in the material even when the stress is reduced to zero. Moreover, the plastic deformation corresponds to a complex process within the material involving the motion of dislocations rather than a simple stretching and compression of the inter-atomic bonds. This motion of dislocations under a given stress is a self-limiting one where the

generation of a relatively high density of the dislocations in a given region of the material impedes their motion. The simple linear relationship between stress and strain thereby gets modified to a more complex one.

As the stress in the material is made to increase to a high value, the material *breaks* or suffers a *fracture* (point B of the stress-strain curve), where the mechanism of fracture once again involves the motion of imperfections or irregularities in the crystalline material. However, I will not enter here into detailed considerations relating to plastic deformation or fracture production, concentrating instead on the linear stress-strain relationship within the elastic limit.

The theory of dislocations was initiated in the year 1934, separately but simultaneously by Taylor, Orowan, and Polanyi - an event that constitutes a landmark in the theory and practice of plastic deformations and flows. Plasticity is a subject with a great many dimensions to it, with immense theoretical and practical relevance.

6.6 Stress-strain relations: elastic constants

While I have introduced the concepts of stress and strain in a deformable body by referring to 'small' volume elements and regions in it, these elements are necessarily of a *macroscopic* dimension, involving a large number of *microscopic* constituents, i.e., atoms or molecules. In this sense, the concept of a vanishingly small element is an idealization, involving an extrapolation down to a very small size where, at the same time, one assumes that the element under consideration continues to be a macroscopic one.

In reality, however, the stress and strain components, while being macroscopically defined quantities, are determined by *microscopic* features like the dispositions of the atoms and molecules, and their interaction forces. These microscopic features serve as a link between the strain and the stress parameters which are thus related in a definite manner for a given material within the elastic limit. It is this relationship that is depicted by the portion from O to A in the graph of fig. 6-8. A complete description of

this relation, even assuming that it is a linear one within the elastic limit, is not a simple matter since it is in reality a linear relationship between two sets of six parameters each. However, things appear simpler if one restricts to *isotropic* materials, i.e., ones in which all directions in space are equivalent. In the following I introduce, for such a material, four elastic constants (also referred to as elastic moduli), of which only two will be later seen to be independent. Each of these is defined with reference to one of a few simple situations involving strain and stress in the material. It is convenient to assume that the body under consideration is a *homogeneous* one when it is no longer necessary to refer to any particular point in it, since the elastic properties of such a body are the same everywhere throughout the body.

The stress-strain relations are, in principle, obtained from the *thermodynamic free energy function* per unit volume of a material, which is defined in terms of the strain parameters and the temperature.

6.6.1 Young's modulus

Imagine a situation where only a tensile stress, say s , is developed in a body along a given direction, as in the case of the wire in fig. 6-3. Let the tensile strain in that direction be ϵ . Experimental observations tell us that within the elastic limit, the two are proportional to each other, i.e., in other words

$$\frac{s}{\epsilon} = \text{constant.} \quad (6-12)$$

The value of this constant depends on the material of the wire, and is termed its *Young's modulus*. The unit of Young's modulus in the SI system is $\text{N}\cdot\text{m}^{-2}$. It is commonly denoted by the symbol Y .

1. All the strain- and stress parameters have *signs* associated with them. For instance, a tensile strain is taken to be in the sense of an elongational one. In the case of an extension, i.e., positive elongation, the sign is positive while for a compression, i.e., negative elongation, the sign of the tensile strain is negative. Similarly, in the case of a tensile stress, if the internal force exerted on a part A of

a body by another part B through the common interface is directed *from A to B*, then the sign of the stress is positive while, for an internal force in the opposite direction the sign of the stress is negative. Shearing strains and stresses also carry their own signs.

2. At times, one refers to the *proportionality limit* in the context of a linear stress-strain relation since elastic deformation does not always correspond to strict proportionality between stress and strain. Experimentally, the proportionality holds only up to a point on the stress-strain curve that may precede the elastic limit. In the following, however, I will not distinguish between the elastic and the proportionality limits.

Problem 6-6

A uniform beam of weight $W = 100 \text{ N}$ rests horizontally on two columns, each of cross-section $\alpha = 0.005 \text{ m}^2$, the Young's modulus of the material of the columns being $Y = 2.0 \times 10^{11} \text{ N}\cdot\text{m}^{-2}$. The center of mass of the beam is at $x = 0 \text{ m}$, where the x-axis is along the length of the beam, while the points of support on the columns are at $x_1 = -0.4\text{m}$ and $x_2 = 0.6\text{m}$ respectively. Find the stresses developed in the columns, and the corresponding longitudinal strains.

Answer to Problem 6-6

HINT: If the forces of reaction exerted by the columns on the beam be W_1 and W_2 , then the conditions of equilibrium of the beam are $W_1 + W_2 = W$, (zero net force), and $W_1x_1 + W_2x_2 = 0$ (zero net moment about center of mass), giving $W_1 = 60 \text{ N}$, $W_2 = 40 \text{ N}$. These are also the forces that the beam exerts on the two columns, in which the stresses developed are $S_1 = \frac{W_1}{\alpha} = 1.2 \times 10^4 \text{ N}\cdot\text{m}^{-2}$, and $S_2 = \frac{W_2}{\alpha} = 0.8 \times 10^4 \text{ N}\cdot\text{m}^{-2}$. The longitudinal strains are $\epsilon_1 = \frac{S_1}{Y} = 6.0 \times 10^{-8}$ and $\epsilon_2 = \frac{S_2}{Y} = 4.0 \times 10^{-8}$ respectively.

Problem 6-7

A metal ring of radius $r = 0.99 \text{ m}$ and area of cross section $\alpha = 10^{-4} \text{ m}^2$ is heated and then mounted on a wheel of radius $R = 1.0 \text{ m}$. On cooling, the ring fits tightly on the wheel. If the Young's modulus of the material of the ring be $2.0 \times 10^{11} \text{ N}\cdot\text{m}^{-2}$, find the inward force per unit length exerted by the ring on the wheel.

Answer to Problem 6-7

HINT: Considering a small element of the ring of length δl after it is mounted on the wheel, the forces on this element, acting at its two ends are $F = S\alpha$ each along the respective outward drawn tangents (fig. 6-11), where S is the elongational thermal stress (see section 8.18) developed in the ring. The resolved parts of these forces perpendicular to PO (where P is the mid-point of the element under consideration and O is the center of the wheel) cancel each other while the resolved parts along PO add up to $2F \sin(\frac{\delta\theta}{2}) = F\delta\theta = F\frac{\delta l}{R}$ (for vanishingly small δl ; check this out). Thus, the inward force per unit length is $f = \frac{F}{R} = \frac{S\alpha}{R}$. The stress S is given by $S = Y \frac{(2\pi R - 2\pi r)}{2\pi r}$. Making use of given values, one gets $f \approx 2.0 \times 10^5 \text{ N}\cdot\text{m}^{-1}$.

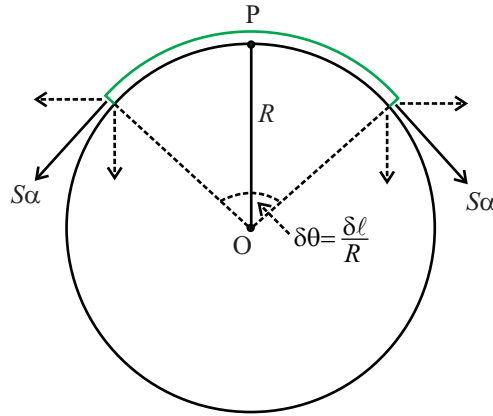


Figure 6-11: Mounting of a heated metal ring on a wheel of radius R ; a small element of length δl of the ring is shown (not to scale); on cooling, a tensile thermal stress is developed in the ring; if S denotes the stress and α the area of cross section of the ring, then the force of tension at either end of the element of length (centered at P) is $S\alpha$; the resolved parts of the forces along the tangent to P cancel while the resolved parts towards the center O add up, giving rise to an inward pressure on the wheel.

6.6.2 Poisson's ratio

Supposing that the state of stress at any given point in a deformable body involves only a tensile stress along a given direction, the corresponding strain is *not*, in general, simply a tensile strain in that direction. Instead, assuming that the given direction is along the x-axis, it is found that a tensile strain is also developed along any direction perpendicular to it, say, along the y- or the z-axis. The strain along the x-axis (say, ϵ)

in this case is termed a *longitudinal* strain while that along the y- or z-axis (say, ϵ') is called a *lateral* one.

For most materials, the signs of the longitudinal and lateral strains are opposite, i.e., for instance, the lateral strain is a compressional one if the longitudinal strain is elongational. Within the proportionality limit the two are proportional to each other, and one has

$$\frac{\text{compressional lateral strain}}{\text{elongational longitudinal strain}} = -\frac{\epsilon'}{\epsilon} = \text{constant}, \quad (6-13)$$

where the negative sign is put in due to the fact that the two strains are usually of opposite signs. The proportionality between lateral and longitudinal strains can be interpreted as a proportionality between longitudinal stress and lateral strain - another instance of a linear stress-strain relationship. The proportionality constant in eq. (6-13), referred to as the *Poisson's ratio* of the material under consideration, is commonly denoted by the symbol σ . A theoretical analysis shows that the value of σ has to lie in the range

$$-1 < \sigma < 0.5. \quad (6-14)$$

For most materials, σ is positive, corresponding to the lateral and longitudinal strains being of opposite signs. A number of artificially produced polymers have been found to be characterized by a negative value of the Poisson's ratio. The inter-atomic bonds in these polymers have a hinge-like structure, where the hinges open up, and adjacent fibres push one another out when a longitudinal elongation is developed, resulting in a lateral elongation (see fig. 6-12).

Problem 6-8

A homogeneous heavy column of length $l_0 = 1.0$ m, area of cross-section $\alpha = 0.01$ m², and weight $W = 50.0$ N is suspended from a fixed point and hangs under its own weight. What is the tensile stress at a point halfway down the length of the column? If the Young's modulus of the material

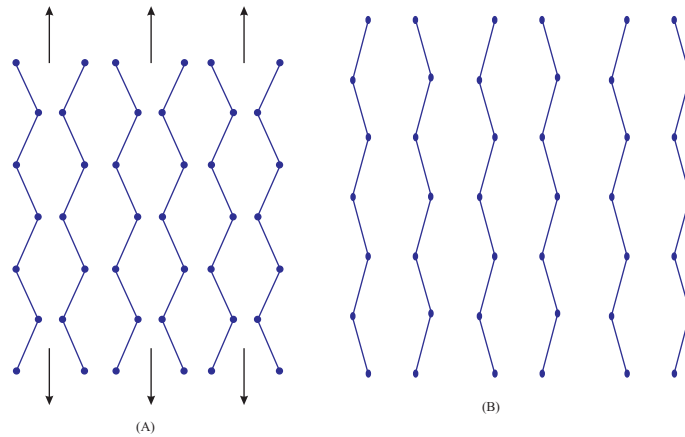


Figure 6-12: Illustrating the origin of a negative Poisson's ratio; the inter-atomic bonds have a hinge-like structure as in (A), which open up when an elongational strain is developed (arrows in (A)), resulting in a lateral expansion, as in (B); while the molecular fibers with a zig-zag structure accommodate one another like intertwined coils in (A), they push away from their neighboring fibers when the hinges open up as in (B).

of the column be $Y = 2.0 \times 10^{11} \text{ N}\cdot\text{m}^{-2}$, and the Poisson's ratio be $\sigma = 0.2$, what is the tensile strain along the length of the column at the point mentioned above and the fractional change in the area of cross-section at that point? What is the stress at a distance $\frac{l_0}{3}$ from the fixed end?

Answer to Problem 6-8

HINT: Consider a cross-section of the column at $x = \frac{l_0}{2}$ (x-axis along the length of the column pointing downward, origin at the point of suspension), and the portion of the column below this cross-section. The weight $\frac{W}{2}$ of this portion is balanced by the stress force exerted by the upper portion of the column. Thus the stress at $x = \frac{l_0}{2}$ is $S = \frac{W}{2\alpha}$, and the strain is $\epsilon = \frac{S}{Y} = \frac{W}{2\alpha Y} = \frac{50.0}{2 \times 0.01 \times 2.0 \times 10^{11}}$. The linear dimension along any direction in the cross-section contracts by a factor $1 - \sigma\epsilon$, and hence the area of cross section becomes $\alpha(1 - \sigma\epsilon)^2 \approx \alpha(1 - 2\sigma\epsilon)$. At the point $x = \frac{l_0}{3}$, the stress is $\frac{2W}{3\alpha}$.

Problem 6-9

A small cube with edge length $a = 0.1 \text{ m}$ and with edges parallel to the axes of a rectangular co-ordinate system is deformed by applying a elongational tensile force of $F_x = 5.0 \text{ N}$ on each of the two faces perpendicular to the x-axis and a compressional tensile force of $F_y = 4.0 \text{ N}$ on each of the two faces perpendicular to the y-axis. If the Young's modulus of the material of the cube be

$Y = 2.0 \times 10^{11} \text{ N}\cdot\text{m}^{-2}$, and the Poisson's ratio be $\sigma = 0.1$, find the changed dimensions due to the forces .

Answer to Problem 6-9

HINT: The stress along the x-axis due to the forces F_x is $S_x = \frac{F_x}{\alpha}$, where $\alpha = a^2 = 0.01 \text{ m}^2$. The resulting strain along the x-axis is $\epsilon_x^{(1)} = \frac{S_x}{Y} = \frac{5.0}{0.01 \times 2 \times 10^{11}}$. This longitudinal strain along the x-axis produces lateral strains along the y- and z-axes, given by $\epsilon_y^{(1)} = \epsilon_z^{(1)} = -\sigma\epsilon_x^{(1)} = -0.1\epsilon_x^{(1)}$. Similarly, the tensile forces F_y produce a tensile strain $\epsilon_y^{(2)} = \frac{4.0}{0.01 \times 2 \times 10^{11}}$ along the y-axis, and corresponding lateral strains $\epsilon_z^{(2)} = \epsilon_x^{(2)} = -\sigma\epsilon_y^{(2)} = -0.1\epsilon_y^{(2)}$ along the z- and x-axes. By the principle of superposition (see sec. 6.6.5), the resultant strains along the x-, y- and z-axes are $\epsilon_x = \epsilon_x^{(1)} + \epsilon_x^{(2)} = 2.3 \times 10^{-9}$, $\epsilon_y = \epsilon_y^{(1)} + \epsilon_y^{(2)} = 1.75 \times 10^{-9}$, and $\epsilon_z = \epsilon_z^{(1)} + \epsilon_z^{(2)} = -0.45 \times 10^{-9}$. The changed dimensions of the cube are then $a_x = a(1 + \epsilon_x)$, $a_y = a(1 + \epsilon_y)$, and $a_z = a(1 + \epsilon_z)$. Now substitute appropriate values.

Problem 6-10

A homogeneous heavy column of cross-sectional area α and length l is suspended from a rigid beam and has a weight W attached to its lower end. The tensile strain at a point at distance $\frac{l}{3}$ from the upper end (see fig. 6-13) is ϵ . If the Young's modulus of the material of the column be Y , find the weight w of the column.

Answer to Problem 6-10

HINT: Consider first the equilibrium of the entire column. The attached weight W exerts a downward pull on it of magnitude W , while the downward gravitational pull, acting through the center of gravity, is w , the weight of the column. Hence there must be an upward reaction force of magnitude $W + w$ acting on it, exerted by the beam (B) at the upper end. Now consider the equilibrium of the upper part of the column, above the cross section at A at a distance $\frac{l}{3}$ from the upper end (refer to fig. 6-13). The forces acting on this part are the reaction force $W + w$ acting at the upper end, the weight $\frac{w}{3}$ acting through the center of mass of the upper part, and the force of tensile stress, exerted by the lower part, acting at the cross section at A. This gives the tensile stress $S = \frac{1}{\alpha}(W + \frac{2w}{3})$, corresponding to which the strain is $\epsilon = \frac{S}{Y}$; one therefore obtains $w = \frac{3}{2}(\alpha\epsilon Y - W)$.

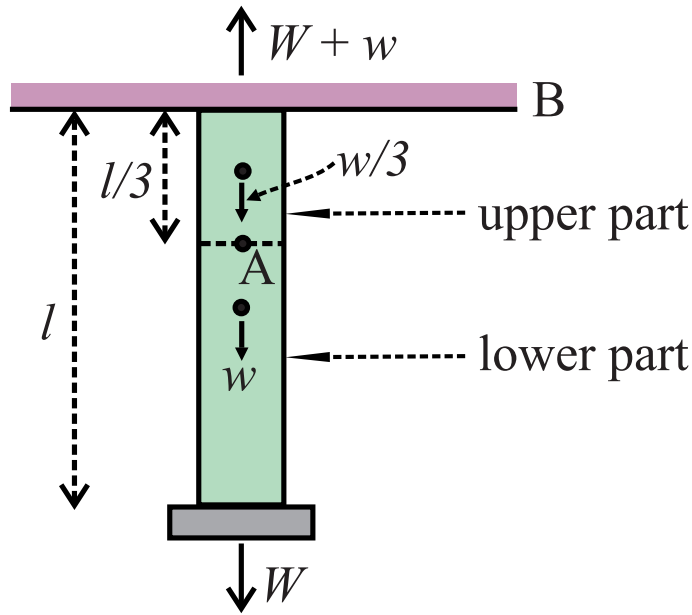


Figure 6-13: A heavy column suspended from a rigid beam, having a weight W attached to its lower end; the tensile stress and strain vary along the length of the column; given the cross-sectional area (α) of the column, the Young's modulus (Y) of its material, and the strain (ϵ) at any given point A (at a distance $\frac{l}{3}$ in the present instance), the weight (w) of the column can be worked out as in problem 6-10.

6.6.3 Modulus of rigidity

The linear relationship between stress and strain within the elastic limit, being a general feature of the elastic behavior of materials, holds for the special case of a pure shear as well. Considering, for instance, a pure shear stress s in the x-y plane, which means that all the other stress parameters at the point under consideration are zero, let us denote the corresponding shearing strain in the x-y plane by ϵ . Then the ratio $\frac{s}{\epsilon}$ is a constant for the material, and is termed its *modulus of rigidity*, or shear modulus. It is commonly denoted by the symbol η .

Fig. 6-14 illustrates the disposition of stress forces for the development of a pure shear in the x-y plane at a point O in a material. Considering a small cube centered at O, a shearing force, say, F acts on the face ABCD perpendicular to the x-axis, along the y-direction. Another force F acts on the face A'B'C'D' along the negative direction of the y-axis so as to balance this force and prevent a rigid translation of the cube. A similar

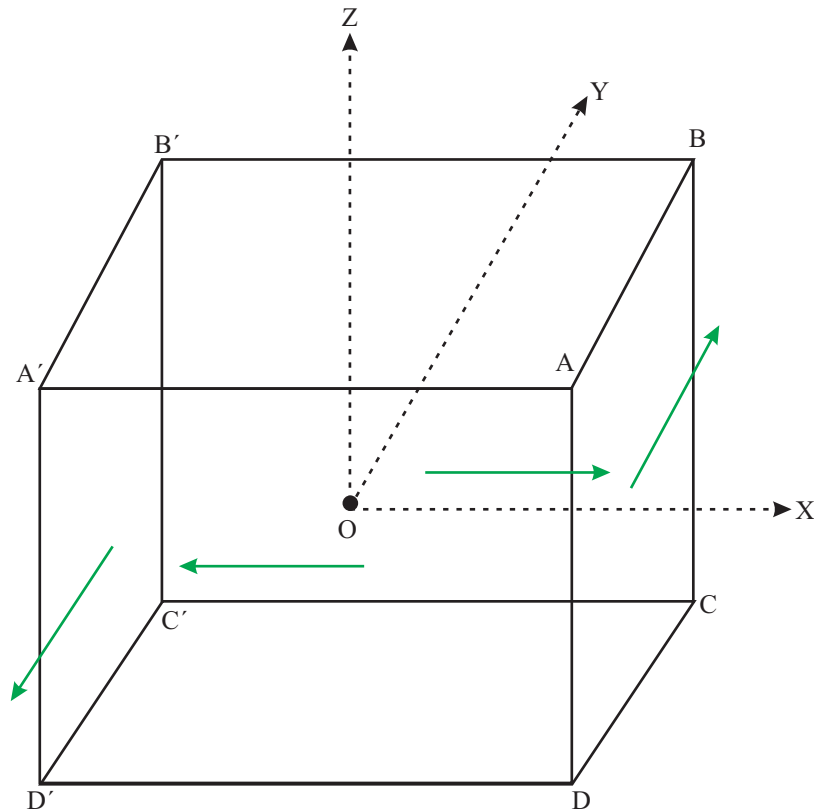


Figure 6-14: Illustrating the shear forces for the development of a pure shear in the x-y plane at the point O in a deformable body; a small cube is imagined around O, with edges parallel to the co-ordinate axes; a pair of shearing forces (each of magnitude F) acts on the two faces perpendicular to the x-axis while another similar pair acts on the faces perpendicular to the y-axis; the cube thereby gets deformed to a parallelepiped with any section parallel to the x-y plane being converted into a rhombus.

pair of forces have to act on the faces perpendicular to the y-axis as shown in the figure so as to balance the moment of the former pair of forces about, say, O, and to prevent a rigid rotation of the cube. As a result of these four shearing forces, the face C'CDD' (or any other section of the cube parallel to the x-y plane) gets deformed to a rhombus (not shown in the figure), corresponding to a pure shear at O.

Problem 6-11

A shearing force $F = 5.0 \text{ N}$ acts on each of the four faces, perpendicular to the x- and y-axes, of a cube of edge length $a = 0.1 \text{ m}$ as shown in fig. 6-14. If the modulus of rigidity of the material of the cube be $\eta = 8.0 \times 10^{10} \text{ N}\cdot\text{m}^{-2}$, find the shear strain developed in the x-y plane. What will be

the angles of the rhombus that the x-y cross-section of the cube gets deformed to?

Answer to Problem 6-11

HINT: The shear stress in the x-y plane is $S_{xy} = \frac{F}{a^2} = \frac{5.0}{0.01} \text{ N}\cdot\text{m}^{-2}$. Hence the shear strain developed will be $\gamma = \frac{S_{xy}}{\eta} = \frac{5.0}{0.01 \times 8.0 \times 10^{10}} = 6.25 \times 10^{-9}$. The angles of the rhombus will be $\frac{\pi}{2} - \gamma$, and $\frac{\pi}{2} + \gamma$.

6.6.4 Bulk modulus

In a similar manner, imagine that a volume stress has developed at some point within a deformable body as a result of longitudinal stresses of identical values along the three co-ordinate axes. There then corresponds a volume strain as well, with identical longitudinal strains along the three axes. The ratio of the two, which is once again a constant for the material under consideration within the elastic limit, is termed its *bulk modulus*, and is commonly denoted by the symbol K :

$$\frac{\text{volume stress}}{\text{volume strain}} = K. \quad (6-15)$$

6.6.5 Principle of superposition

The linear relation between the strain and stress parameters within the elastic limit is consistent with the *principle of superposition*. Thus, let the state of strain at a point be described by the strain parameters $\epsilon_i^{(1)}$ ($i = 1, 2, \dots, 6$), corresponding to stress parameters $S_i^{(1)}$ ($i = 1, 2, \dots, 6$). Let us now consider another state of strain, described by strain parameters $\epsilon_i^{(2)}$, corresponding to stress parameters $S_i^{(2)}$. Then, imagining a state of strain described by strain parameters $\epsilon_i^{(1)} + \epsilon_i^{(2)}$, the corresponding stress parameters will be $S_i^{(1)} + S_i^{(2)}$ ($i = 1, 2, \dots, 6$).

The principle of superposition can be stated in an alternative form, in terms of forces applied to a deformable body, and deformations produced by these forces.

Suppose that a deformable body is subjected to a system of external forces, say, X_1, X_2, \dots, X_n , in response to which stresses and strains are generated in it, resulting in relative displacements between various parts of the body. Suppose the displacement of one such

part, measured at a chosen point on the body, is u , where the displacement may be a linear or angular one. Assuming that the deformation of the body is within the elastic limit, the linear relationship between the stresses and strains implies a correspondingly linear relation between the applied forces and the resulting displacements. Thus, if a second set of forces, say, Y_1, Y_2, \dots, Y_n produces a displacement v , then a set of forces $X_1 + Y_1, X_2 + Y_2, \dots, X_n + Y_n$ will cause a displacement $u + v$.

As an example, fig. 6-15 shows a *cantilever*, which is a rod rigidly clamped at one end such that it can bend under a load applied to the free end or to any other point on the rod. Assuming for the sake of simplicity that the rod is weightless, suppose the depression of the free end is δ_1 when a load W_1 is applied at the point A of the rod and again, when a load W_2 is applied at the point B (with the load at A removed), let the depression of the free end be δ_2 . Then the principle of superposition implies that when loads W_1 at A and W_2 at B are applied simultaneously, the depression at the free end will be $\delta_1 + \delta_2$.

If the rod be a heavy one then there occurs a depression due to its own weight. In that case, δ_1 and δ_2 are to be taken as the depressions caused by W_1 and W_2 respectively *in addition to* the depression caused by the weight of the cantilever itself

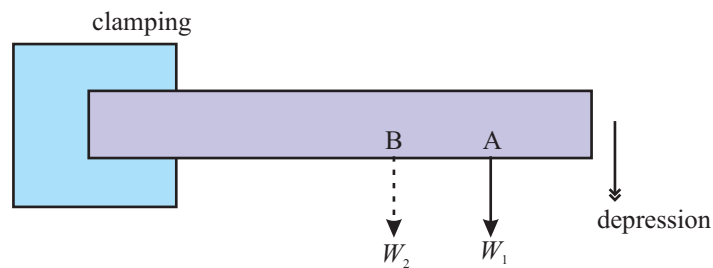


Figure 6-15: A cantilever consisting of a light rod clamped at one end; loads W_1 and W_2 , applied one in the absence of the other at A and B respectively, produce depressions δ_1 and δ_2 at the free end of the rod; if the loads are applied simultaneously, the depression at the free end, according to the principle of superposition, will be $\delta_1 + \delta_2$.

6.6.6 Relations between the elastic constants

In the present section we have met with the definitions of Young's modulus (Y), Poisson's ratio (σ) modulus of rigidity (η), and bulk modulus (K) for an isotropic material. If, moreover, the material under consideration is homogeneous, then the values of these coefficients are the same everywhere in it while, for an inhomogeneous material, the coefficients differ from point to point.

However, all these four quantities are not independent. For any given isotropic material, only *two* of these are independent while the values of the other two can be determined from these. Taking, for the sake of concreteness, the bulk modulus (K) and the modulus of rigidity (η) as the two independent elastic constants of a material, its Young's modulus (Y) and the Poisson's ratio (σ) are given by the formulae

$$Y = \frac{9K\eta}{3K + \eta}, \quad (6-16a)$$

$$\sigma = \frac{3K - 2\eta}{2(3K + \eta)}. \quad (6-16b)$$

In this context, refer to problem 6-12 below for the derivation of the relation between the constants Y , K , and σ . Any such derivation involves the assumption of some particular stress configuration at any given point in a deformable body. The end result, however, is independent of the assumed stress configuration. One such configuration will be considered in problem 6-12.

Crystalline materials are, in general, *anisotropic*, i.e., all the directions in space are not equivalent within such a material. The linear relation, valid within the elastic limit, between the stress and strain parameters for such an anisotropic crystalline material involves more than two independent elastic constants, where the actual number of constants depends on the three dimensional structure of the crystal under consideration.

However, one often finds that a material with an underlying crystal structure, is *effectively* isotropic. This is because the material under consideration is made up of a large number of small crystalline constituents, where all these small crystals are oriented randomly in the material. While each small crystal in itself is anisotropic, the crystals oriented randomly in the vicinity of any given point in the material have the effect of making all physical properties appear the same, on an average, in all directions around the point. The elastic properties for such a material can be described completely in terms of just two elastic constants in a manner similar to the case of an isotropic material.

6.6.7 Elastic properties of fluids

Liquids and gases are together referred to as *fluids*. A fluid differs from a solid in the important respect that it has no *rigidity*. This actually means that a shearing strain *cannot* develop in a fluid at rest. Let me explain this with reference to the definition of shearing strain we are already familiar with.

Consider a small volume element in the form of a rectangular parallelepiped around any given point, say O, in a fluid at rest.

A rectangular parallelepiped differs from the cube I referred to earlier in defining the strain parameters, but one can equally well start from a rectangular parallelepiped which has its edge lengths different from one another (one can denote these as, say, a , b , and c) but, at the same time, has all its angles measuring $\frac{\pi}{2}$. The difference in edge lengths introduces a slight modification in the way the tensile strain parameters are introduced in that the changed edge lengths due to a strain are, respectively, $a(1 + e_1)$, $b(1 + e_2)$, $c(1 + e_3)$, but the changed *angles* in terms of which the shearing strain parameters are introduced, remain the *same* as in the case of the cube.

Now imagine a strain to have developed in the fluid. It will be found that, in spite of the strain, the rectangular parallelepiped *retains its shape*, i.e., is transformed to

another rectangular parallelepiped, with some or all of the edge lengths altered from their previous values. In other words, all the angles between adjacent sides of the volume element continue to be $\frac{\pi}{2}$, and all the shear parameters $\gamma_1, \gamma_2, \gamma_3$ introduced earlier are zero.

Moreover, when one considers the internal forces exerted on the fluid in the volume element under consideration, acting through its boundary surfaces, one finds that all these forces act along the *normals* to these boundary surfaces, toward the interior of the volume element. In other words, *no shear stress can develop in a fluid at rest*. Here I have included the clause ‘at rest’ because a fluid in motion can, in general, accommodate a shear stress. Indeed, if a shear stress is made to develop in a fluid, its equilibrium will necessarily be disturbed and a *flow* will be created in it in order that the accompanying shearing *strain* is released.

Another important feature of the internal forces in a fluid is that the magnitude of the force per unit area, acting on the fluid contained in the given volume element having the shape of a rectangular parallelepiped, through each of its boundary surfaces is the *same* for all the six surfaces. As mentioned above, the direction of the force is inward for all the surfaces. In other words, the only form of stress that can develop in a fluid at rest is a *bulk stress*. Indeed, according to the definition of *pressure* in a fluid, one has,

$$\text{bulk stress in a fluid at a point} = -p, \quad (6-17)$$

where p denotes the pressure at that point.

Recall that, in the definition of tensile stress and bulk stress, the internal forces were taken to be directed *away* from the volume element under consideration.

Thus a fluid differs from an isotropic solid in respect of its elastic constants in that, while *two* independent constants are needed for a complete description of the elastic properties of an isotropic solid, only *one* constant, namely, the bulk modulus (K) suffices for a fluid, the values of the other elastic constants (Y, η, σ) not being relevant. Indeed,

for a fluid, one effectively has,

$$Y = 0, \eta = 0, \sigma = \frac{1}{2}, \quad (6-18)$$

these being consistent with (6-16a), (6-16b).

Problem 6-12

Imagine a small cube of edge length a within a deformable body with its edges parallel to the three axes of a rectangular co-ordinate system. Assume that the only stress force acting on the cube is a tensile force F acting on each of the two faces perpendicular to the x -axis. Find the volume stress and the associated tensile stresses along the three co-ordinate axes at any point (say, at the center) of the cube. If, Y, K, σ be the Young's modulus, bulk modulus, and Poisson's ratio of the material of the body, find the volume strain and the associated tensile strains. Obtain from these the relation between the three elastic constants. What are the altered edge lengths of the cube?

Answer to Problem 6-12

HINT: The general description of the state of stress at a point is in terms of tensile stresses s_x, s_y, s_z and shear stresses s_{xy}, s_{yz}, s_{zx} , of which the shear stresses are all zero in the present problem. Instead of the three tensile stresses, one can describe the state of stress in terms of a volume stress $s = \frac{s_x + s_y + s_z}{3}$, and three *associated* tensile stresses $s'_x = s_x - s, s'_y = s_y - s, s'_z = s_z - s$. In the present problem, $s_x = \frac{F}{a^2}, s_y = 0, s_z = 0$, and hence, $s = \frac{s_x}{3} = \frac{F}{3a^2}, s'_x = \frac{2s_x}{3} = \frac{2F}{3a^2}, s'_y = -\frac{s_x}{3} = -\frac{F}{3a^2}, s'_z = -\frac{s_x}{3} = -\frac{F}{3a^2}$.

The volume strain will then be $\epsilon = \frac{s}{K} = \frac{F}{3a^2 K}$, while the associated tensile strains are obtained by considering s'_x, s'_y, s'_z separately and then applying the principle of superposition where, in each case, we have to take into account the strains produced in perpendicular directions. In other words, the associated tensile strains are $\epsilon'_x = \frac{1}{Y}(s'_x - \sigma(s'_y + s'_z)), \epsilon'_y = \frac{1}{Y}(s'_y - \sigma(s'_z + s'_x)),$ and $\epsilon'_z = \frac{1}{Y}(s'_z - \sigma(s'_x + s'_y))$ where, in these expressions, s'_x, s'_y, s'_z are to be replaced with their values obtained above.

Noting that the volume strain ϵ involves tensile strains $\epsilon''_x = \epsilon''_y = \epsilon''_z = \frac{\epsilon}{3}$ along the three axes, the resultant tensile strains along the three axes are seen to be $\epsilon_x = \epsilon'_x + \epsilon''_x = \frac{1}{Y}(s'_x - \sigma(s'_y + s'_z)) + \frac{s}{3K} =$

$\frac{1}{Y}(s'_x - \sigma(s'_y + s'_z)) + \frac{s_x}{9K}$, and corresponding expressions for ϵ_y, ϵ_z obtained similarly where, in these three expressions, we can substitute for s'_x, s'_y, s'_z the expressions obtained above. The tensile strains can also be obtained directly in terms of s_x, s_y, s_z as $\epsilon_x = \frac{1}{Y}(s_x - \sigma(s_y + s_z)) = \frac{F}{a^2 Y}$, and analogous expressions for ϵ_y, ϵ_z .

On comparing the expressions for $\epsilon_x, \epsilon_y, \epsilon_z$ obtained in these two approaches, one obtains, in the case of ϵ_x , for instance,

$$\frac{F}{a^2 Y} = \frac{1}{Y(s'_x - \sigma(s'_y + s'_z)) + \frac{\epsilon}{3}} = \frac{1}{Y} \left(\frac{2F}{3a^2} + \frac{2\sigma F}{3a^2} + \frac{F}{9a^2 K} \right), \text{ i.e., } Y = 3K(1 - 2\sigma)$$

. Similar comparisons for ϵ_y, ϵ_z give identical results.

This relation between the three elastic constants is consistent with the formulae (6-16a) and (6-16b).

6.7 Strain energy

Consider once again the loading of a weightless wire vertically suspended from one end (fig. 6-3). Suppose that the loading is done in a succession of infinitesimally small steps so that no energy is lost to the vibrational modes of the atoms and molecules making up the wire. The work performed on the wire in this process of slow ('quasi-static') loading then remains stored in the wire itself, and is termed the *strain energy* of the wire. Imagining the bonds between neighboring atoms of the material of the wire as so many springs, the strain energy is the energy stored in the springs in extending or compressing these during the straining process. This energy of the springs can be recovered if these are slowly brought back to their original lengths, i.e., in other words, if the loading of the wire is released slowly.

It is not difficult to work out the energy supplied to the wire in the form of work during the process of slow loading. Suppose that, at any intermediate stage of the loading process, the load is W' , for which the length of the wire is l' . As the load is increased by a small amount to, say, $W' + \delta W'$, the length increases by $\delta l' = l' \frac{\delta W'}{AY}$, where A stands for the area of cross-section of the wire and Y for the Young's modulus of the material

of the wire (reason this out; hint: increment in strain = $\frac{1}{Y} \times$ increment in stress).

The work done in this small step equals the force (=load) times the increment in length, i.e., $W' \delta l' = \frac{l}{AY} W' \delta W'$. One has now to add up the amounts of work done in all these small steps of loading to arrive at an expression for the strain energy. Going over to the limit of infinitesimally small increments of loading, the sum is seen to convert to an integral, whereby the expression for the strain energy works out to

$$\Delta E = \frac{l}{AY} \int_0^W W' dW' = \frac{lW^2}{2AY}. \quad (6-19a)$$

Here we have not considered the change in the length l in evaluating the integral, treating it effectively as a constant, because the change in l due to the loading alters the integral only marginally.

A quantity of considerable relevance is the *strain energy per unit volume* or the strain energy *density* of the wire. Expressing the volume (V) as $V = Al$, we obtain,

$$\frac{\Delta E}{V} = \frac{W^2}{2A^2Y}, \quad (6-19b)$$

or, equivalently,

$$\text{strain energy per unit volume} = \frac{1}{2} \text{stress} \times \text{strain}, \quad (6-19c)$$

where the expressions for stress ($= \frac{W}{A}$) and strain ($= \frac{\text{stress}}{Y}$) in the wire for the load W have been made use of. Within the elastic limit, the strain energy for a given load W is a thermodynamic variable (see chapter 8 for an introduction to basic thermodynamic concepts) that depends only on the states of stress and strain for that load, and not on the manner the loading has been done, providing only that it is done quasi-statically.

An interesting question that may come up here is, what is likely to happen if the loading in the wire in the above example is *not* done slowly? For an answer, think of the following similar question: what happens if a spring is made to expand or contract all of a sudden? If, for instance, the spring is loaded suddenly, it does not reach its

extended state in a monotonic manner, but keeps on oscillating about its extended configuration where the oscillations are gradually damped. Thus, the energy supplied to the spring in the form of work does not all go to increase its potential energy of extension since part of the energy supplied appears as its kinetic energy of oscillation. The latter gets dissipated in the material of the spring as the oscillations are damped out, eventually raising the temperature of the material of the spring.

An exactly similar thing happens for a sudden straining of the wire within the elastic limit. The energy supplied to the wire in the form of work done in extending it is $W\Delta l$ since now the load W remains constant throughout the extension (because the process of extension occurs without the load being adjusted slowly), the amount of extension under the load W being $\Delta l = \frac{lW}{AY}$. In other words, the energy supplied in the form of work is *twice* the strain energy (check this out; refer to eq. (6-19a)). Half of this gets stored in the wire in the form of strain energy while the other half gets dissipated in the body, raising its temperature.

In this derivation the expressions for stress and strain have been arrived at by referring to the wire as a whole because of the strain being uniform throughout the wire (a consequence of the wire having been assumed weightless). For an inhomogeneously strained body one has to make the derivation separately for each small volume element, where the stress and strain in that element will feature, and then add up the strain energies for all these elements to arrive at the expression for the total strain energy of the body. It is not difficult to carry out such a derivation for a heavy wire or a rod. A quantity more relevant than the total strain energy is the ratio $\frac{\delta\Delta E}{\delta V}$, where one considers a small volume element δV around any given point O in the body (for which the strain energy is $\delta\Delta E$). This is referred to as the *strain energy density* at the point O and, on going through the exercise, one ends up once again with the expression in eq. (6-19c),

$$\frac{\delta\Delta E}{\delta V} = \frac{1}{2} \text{stress} \times \text{strain}, \quad (6-20)$$

where, in this expression, one has to interpret the stress and strain as those pertaining to the point O under consideration.

A similar derivation works for *other* types of straining in the body under consideration. For instance, in the case of a pure shear one again obtains an expression of the form (6-20) where now the stress and strain stand for the respective quantities under shear. In the case of a pure bulk strain one has to use the bulk stress and the bulk strain in the expression for the strain energy density. In the case of a mixed strain one can express the strain energy density at any given point as a sum of expressions of the form (6-20), where the expressions in the sum involve the principal components of stress and the principal components of strain at that point (more generally, one will have a sum involving the products of components of the stress tensor and those of the strain tensor). However, the expression for the strain energy density may still be a relatively complex one because of the fact that the principal axes of strain may not be the same as the principal axes of stress.

Problem 6-13

A light wire of length $l_0 = 1.0$ m and area of cross-section $\alpha = 1.0 \times 10^{-6}$ m² has a load of weight $W = 10$ N attached at one end, while the other end is attached to a fixed point from which the wire hangs vertically. If the Young's modulus of the material of the wire be $Y = 2.0 \times 10^{11}$ N·m⁻², find its strain energy.

Answer to Problem 6-13

HINT: The tensile stress developed in the wire is $S = \frac{W}{\alpha}$, and the corresponding tensile strain is $\epsilon = \frac{S}{Y} = \frac{W}{\alpha Y}$. The strain energy per unit volume is $\frac{1}{2} \frac{W}{\alpha} \frac{W}{\alpha Y}$. Ignoring the change in volume of the wire due to the deformation, the strain energy is $\frac{1}{2} \frac{W}{\alpha} \frac{W}{\alpha Y} l_0 \alpha = \frac{1}{2} \frac{W^2 l_0}{\alpha Y} = \frac{100 \times 1.0}{2 \times 10^{-6} \times 2 \times 10^{11}}$ J, i.e., 2.5×10^{-4} J.

NOTE: Assuming that the load W is attached to the wire and then let go, the wire performs damped oscillations before finally attaining equilibrium with the strain worked out above. In this case, the strain energy is half the work done by the force of gravity on the load as it descends due to the strain developed in the wire (i.e., half the amount by which the potential energy of the load

is lowered as it descends to the maximum depth in the course of its oscillation; check this out). The other half goes to increase the thermodynamic internal energy of the wire. This accounts for the energy balance of the system.

Problem 6-14

Two weightless rods A, B, made of materials with Young's moduli Y_1, Y_2 , are of the same cross-section α , and have lengths l_1, l_2 . The two are joined end-to-end and suspended from a rigid horizontal beam as in fig. 6-16, with a weight W attached to the free end. Find the elongation of either rod and the strain energy of the system

Answer to Problem 6-14

HINT: The rods being weightless, the stress and strain are uniform in either of these. The tensile stress at any point in either rod is $\frac{W}{\alpha}$, and hence the strains are, respectively, $e_1 = \frac{W}{\alpha Y_1}$, $e_2 = \frac{W}{\alpha Y_2}$. The corresponding strain energy densities are then $\frac{W^2}{2\alpha^2 Y_1}$, $\frac{W^2}{2\alpha^2 Y_2}$, giving the strain energy of the system as $\mathcal{E} = \frac{W^2}{2\alpha} (\frac{l_1}{Y_1} + \frac{l_2}{Y_2})$. The elongations are $\delta l_1 = e_1 l_1 = \frac{W}{\alpha} \frac{l_1}{Y_1}$ and $\delta l_2 = e_2 l_2 = \frac{W}{\alpha} \frac{l_2}{Y_2}$.

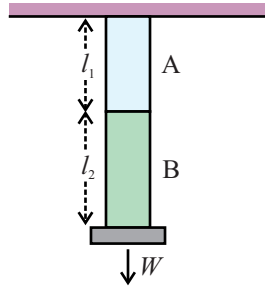


Figure 6-16: Two weightless rods with the same cross-section and made of different materials, connected end-to-end; the combination is suspended from a rigid beam, with a load W attached to the free end; the elongations of the two rods and the strain energy of the system are as worked out in problem 6-14).

Chapter 7

Mechanics of Fluids

7.1 Introduction: the three states of matter

Materials we see around us can be classified into two broad groups on the basis of their states of aggregation, i.e., on the way they are held together: bodies made of *solids* appear to have definite volumes and shapes, while *fluids* do not have definite shapes because they *flow* from one place to another. We have met with this special property of fluids in chapter 6 while looking at elastic properties of matter.

Fluids can be further classified into *liquids* and *gases* which are easily distinguished in our everyday experience: while a given mass of liquid is characterized by a definite volume that can be made to vary only to a relatively small extent, and the liquid can be stored in an open container, a gas has to be kept in a closed container so as to prevent it from escaping, and it fills up the space inside the container so as to attain the volume of the latter.

This difference in behavior among the three states of matter is related, in a manner of speaking, to the average magnitude of the force of interaction between the basic constituents making up the substances, namely, their molecules. In a solid, the molecules are held together in a fixed shape and size by relatively strong attractive forces, and the extent to which individual molecules can move about is small. In a liquid the molecules

can move about quite freely, but each molecule still feels an attractive force due to all the other molecules taken together, so that it cannot escape from the volume occupied by the aggregate of all these molecules. In a gas, on the other hand, the average intermolecular force is so feeble as to exert negligible influence on the motions of individual molecules which tend to fly apart almost freely. All these forces are, however, electromagnetic in origin, and the difference lies principally in the *correlations* among the molecules under these forces. The correlations, along with the average magnitude of the intermolecular forces, decrease with an increase in the average separation between them.

As is well known, these three states of matter are not distinct from one another in an absolute sense, and there occur *changes of state* depending on external circumstances, characterized by *pressure* and *temperature*. We will look into such transitions from one state of matter to another in chapter 8.

Solids, liquids and gases are said to correspond to three distinct *phases* of matter. A phase corresponds to some definite state of aggregation of and correlation among the constituents making up a material. While the term constituents commonly refers to the molecules of the material under consideration, it can, at times, be used to refer to microscopic particles at other levels of description like, for instance, *electrons* moving about through a framework of ions. Depending on the way a large number such microscopic constituents are held together and at the same time move about, making up an aggregate of macroscopic proportions, the latter can be in one of various possible phases. The phases are distinguished from one another through one or more of their physical characteristics like, for instance, the flow properties distinguishing the solids from the fluids.

Looked at this way, the term *phase* has a broader connotation than simply the classification of matter into solids, liquids and gases. For instance, one can, under certain circumstances, distinguish between *conducting* and *superconducting* phases in respect of electrical conductivity properties of a crystalline solid. Similarly it is possible to distinguish between a normal fluid and a *superfluid*, the latter being a phase characterized

by quite exceptional flow properties arising under special circumstances.

Though the term *state* of matter is a loose one compared to the term *phase*, it is convenient when the simple distinction between solids, liquids, and gases is to be referred to. However, here again there may be situations demanding special considerations. For instance, the *plasma* is a gaseous state of matter where a gas is made up of two distinct *components*, the electrons and the ions, that move about under the action of electrical and magnetic forces. An ordinary gas is converted into the plasma state at high temperatures when a large number of its atoms get *ionized*, i.e., when one or more electrons come out of each of these atoms. The plasma state is characterized by special electrical and magnetic properties of the gas that can combine with its flow properties in novel and diverse ways, though the plasma does not constitute a distinct phase in the proper sense of the term. In the present chapter we will be concerned with static and dynamic properties of ordinary fluids. We will be especially interested in liquids, though a number of considerations in this chapter will apply to gases as well. Physical properties of gases will be looked into in chapter 8.

7.2 Fluids in equilibrium

7.2.1 Internal forces in a fluid: pressure

In describing the static and dynamic properties of fluids, a number of relevant physical variables are density, compressibility, surface tension, viscosity, specific heat, and thermal conductivity. However, not all of these are found to be of equal importance in any given situation. The liquids we will be concerned with in this chapter will, unless otherwise stated, be assumed to be *incompressible*, i.e., their density will be assumed to be independent of pressure. In addition, their viscosity will also be disregarded till later sections in this chapter. Further, we will assume that no heat transfer takes place within the fluid during time intervals of relevance. These are idealizations which, nevertheless, are found to yield useful results in numerous situations of interest.

Effects of surface tension and viscosity will be considered separately in later sections of

this chapter. For the time being we will be particularly interested in *pressure* and its variation in a fluid, with special emphasis on pressure in an incompressible liquid.

In order to understand the motion of a fluid and the conditions for its equilibrium, it is necessary to identify the *internal forces* within it. Consider a point P in a fluid and imagine a small area, which can be taken as part of a plane surface, around this point, dividing the fluid into two portions which we designate as A and B in fig. 7-1. These two portions of the fluid will exert forces on one another due to interactions among their microscopic constituents, constituting the *stress* at that point.

1. The *viscosity* of the fluid is of relevance here. Viscosity is nothing but a kind of internal friction in the fluid that tends to prevent relative motion of contiguous parts of it. This property of internal friction finds expression in its *coefficient of viscosity* (refer to sec. 7.5.3 below; actually, a fluid may be characterized by two distinct coefficients of viscosity, of which only one is commonly found to be relevant). However, as mentioned above, we will assume for the time being that these internal forces of friction are absent, an idealization that is useful and convenient in understanding the state of stress in a fluid at rest.
2. While the density of a compressible fluid is related to its pressure, this relation, in turn, depends on the temperature. The relation between density, pressure, and temperature differs from one fluid to another and is further dependent on whether the fluid is in the liquid or in the gaseous phase. One consequence of compressibility is the propagation of *acoustic waves* in the fluid (refer to chapter 9). As mentioned above, our considerations in the present chapter will be mostly confined to liquids, many of which can be described as close approximations to incompressible fluids. Among the various static and dynamic features that will emerge, some may be seen to correspond, in a certain sense, to features of compressible fluids as well, including gases. For instance, the definition of characteristics like pressure and viscosity remain the same, Bernoulli's principle can be expressed in a form applicable to flows of compressible fluids for which dissipative effects can be ignored, forces of viscous drag act on a body moving through a compressible fluid as through an incompressible one, and a number of features of boundary layer flow and turbulence, discussed in later sections in this

chapter, are qualitatively similar for incompressible and compressible fluids. With this in mind, the term 'fluid' will be used in a loose sense in much of our considerations in the following sections, where statements arrived at for incompressible fluids will apply, in a qualitative sense, to compressible ones as well.

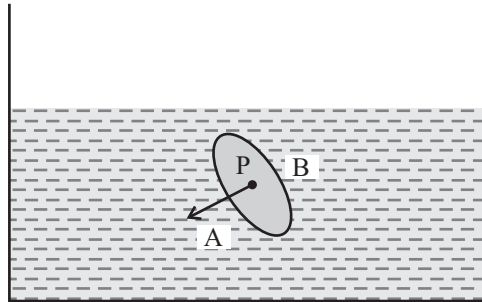


Figure 7-1: Illustrating the idea of pressure in a fluid at any given point (P); a small surface imagined around P divides the body of the fluid into two portions, A and B, where B exerts a force on A along the normal to the imagined surface, directed toward A; this force per unit area of the surface is the pressure at P.

A fundamental property of a fluid is that the internal force exerted by, say, B on A across the surface separating the two parts, acts along the *normal* to this surface whenever the fluid is in equilibrium. This feature of the internal force being normal to the surface dividing the two parts exerting force on one another continues to characterize the internal force even when the fluid is in motion, provided the above assumption of negligible viscosity of the fluid holds. However, if the viscosity of the fluid is taken into account, there appear tangential as well as normal components of the force, though the tangential components disappear when the fluid is at rest, i.e., is in equilibrium, and the normal component is then completely described by pressure (see below). I repeat that the viscosity of the fluid will be assumed to be zero in the present and the following sections so that the internal forces can be assumed to be normal to the dividing surface *both* when the fluid is at rest and is in motion. Effects of viscosity will be considered in section 7.5.

Referring once again to the internal force exerted by B on A through the imagined surface around the point P in fig. 7-1, it is found that this force, which is normal to the

surface, is, moreover directed inward, i.e., toward A, away from B. Further, the magnitude of this normal force per unit area of the dividing surface is found to be *independent* of the orientation of the surface, i.e., it remains the same regardless of which way the normal to the surface points.

This inward normal force per unit area of the surface is termed the *pressure* at the point P. While we have defined the pressure with reference to the force exerted by B on A, it could equally well have been defined in terms of the force exerted by A on B. This force is equal and opposite to the force exerted by B on A, and is directed away from A, into the region occupied by B, on which it acts.

Evidently, the unit of pressure is that of force per unit area, i.e., $\text{N}\cdot\text{m}^{-2}$, also referred to as the *pascal* (Pa).

7.2.2 Pressure in an incompressible liquid

The pressure in a fluid can vary from point to point, depending on the external forces acting on it and the velocities of the fluid particles at these various points. To begin with, we assume the fluid to be in equilibrium. A necessary condition for equilibrium to hold is that the vector sum of external and internal forces on each and every volume element of the fluid has to be zero. Let us apply this condition to an incompressible liquid at rest where the only external force acting on the liquid particles is the force of gravity (not considering, for the time being, the contact forces exerted by the walls of the containing vessel which act only on the elements in contact with the walls). We assume that the acceleration due to gravity is uniform throughout the volume of the liquid, being of magnitude g . Relevant physical parameters like the temperature and the composition of the liquid will also be assumed to be constant throughout the volume, so that the density ρ will similarly be constant.

Figure 7-2 shows a portion of the liquid imagined to be enclosed within a cylindrical surface with horizontal top and bottom faces, each of area, say δS , the cylinder being vertical and of height h . The mass of liquid inside this cylinder, of volume $h\delta S$, is $\rho h\delta S$, and so the external force on this portion of the liquid, acting vertically downward, is

$g\rho h\delta S$. The internal force on this portion, exerted by the surrounding liquid is not difficult to calculate.

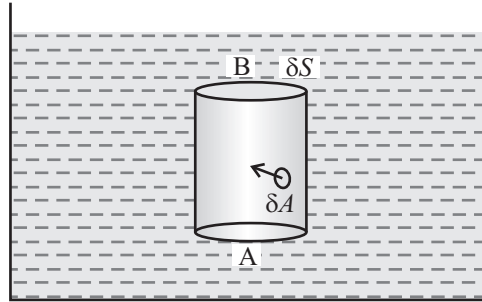


Figure 7-2: Illustrating the derivation of pressure at any point in an incompressible liquid in equilibrium under the action of gravity; a cylindrical volume is imagined, with top and bottom surfaces, each of area δS , horizontal; A and B are any two points on the bottom and top surfaces of the cylinder, at which the pressures are p_1 and p_2 respectively; δA is a small area on the vertical curved surface of the cylinder.

Let p_1 and p_2 be the pressures at any two points (say, A and B) on the bottom and top surfaces of the cylinder respectively. Assuming that the area δS is sufficiently small, the exact location of A or B within the area δS is not of relevance here. According to the definition of pressure, the internal force exerted by the portion of liquid below the bottom surface on the liquid inside the cylindrical volume is $p_1\delta S$ directed vertically upward, while that exerted by the liquid above the top surface is $p_2\delta S$ directed vertically downward. The resultant of these two is a force $(p_2 - p_1)\delta S$ directed downward. Added to this, we have to consider the internal force acting through the vertical surface of the cylinder.

Considering a small element of area, say, δA on this surface, the force exerted on the portion of liquid under consideration by the liquid outside the cylindrical volume through any such surface element is a horizontal one since it has to be normal to the element δA . Hence, considering all such area elements making up the vertical surface of the cylindrical volume, the resultant of all these forces will again be a horizontal one. Let us, for the time being, denote this horizontal force exerted on the volume of liquid under consideration by the surrounding liquid through the vertical surface of the cylinder, as

F.

Thus we have reduced all the forces acting on the portion of liquid under consideration, to three forces - an external force $g\rho h\delta S$ acting vertically downward, an internal or stress force $(p_2 - p_1)\delta S$, again acting vertically downward, and another force **F**, again of internal origin, in the horizontal direction.

Here, the term 'internal' means internal to the liquid as a whole, exerted on the portion of the liquid inside the cylindrical volume, by the liquid outside this volume. While considering the equilibrium of the portion of the liquid inside the cylindrical volume, it is not necessary to consider internal forces exerted by volume elements *within* this portion of the liquid on one another, since such internal forces cancel one another in pairs by virtue of Newton's third law.

The condition of equilibrium of the liquid within the cylindrical volume can then be stated as the requirement that the resultant of the above three forces must be zero. But this means that the horizontal force **F** has to be zero regardless of the other two forces since the latter are vertical and cannot balance the horizontal force. In other words, *the stress force acting on the portion of the liquid under consideration through the vertical surface of the cylinder has to be zero* for equilibrium to hold. The remaining two forces being both directed vertically downward, one must have

$$g\rho h\delta S + (p_2 - p_1)\delta S = 0, \quad (7-1)$$

or, the area δS being small but otherwise arbitrary:

$$p_1 - p_2 = h\rho g. \quad (7-2)$$

This means that, starting from any point within the liquid, as one moves vertically downwards by a distance h , the pressure increases by the amount $h\rho g$. Conversely, if one moves vertically upwards through a height h , the pressure decreases by the same amount.

More generally, considering any two points, not necessarily in the same vertical line, the result (7-2) continues to hold where h represents the separation between the two points along the vertical direction (see fig. 7-3). You will see this as you work out the problem below.

Problem 7-1

Consider two points A and B as shown in fig. 7-3, where A and B are not necessarily in the same vertical line. Show that the pressures p_1 and p_2 at these two points are still related by eq. (7-2), where h stands for the separation between the two points along the vertical direction.

Answer to Problem 7-1

Consider a point C vertically below B at a depth h . Thus, the line joining A and C is a horizontal one. The figure shows a narrow horizontal cylinder with its axis along AC and end planes around A and C, of area δS each where δS is vanishingly small. If p_1 and p'_1 denote the pressures at A and C respectively, then the net force on the liquid contained within the cylinder exerted through the end planes will be $(p_1 - p'_1)\delta S$, acting along the direction from A to C (reason this out). The remaining forces on this portion of the liquid are its weight and the internal (stress) force exerted through the curved surface of the cylinder, which all act in a direction perpendicular to AC. The equilibrium of the portion of the liquid under consideration then requires that the resultant of these latter forces is zero and, in addition, $(p_1 - p'_1)\delta S = 0$, implying $p_1 = p'_1$. Thus, $p_1 - p_2 = p'_1 - p_2 = h\rho g$, where the symbols have meanings as already explained.

7.2.3 Thrust of a fluid

Figure 7-4(A) shows a liquid at rest held in a vessel, where the region of space occupied by the liquid is imagined to be divided into two parts A and B, the two being separated by a common interface C. Looking at any of these two parts, say, the one marked A, the forces that keep this part of the liquid in equilibrium are: (i) external forces like, for instance, the force of gravity acting on the liquid particles, (ii) the force exerted by the walls of the vessel with which A is in contact, and (iii) the force exerted by the portion B of the liquid on the portion A through the interface C. Of these, the last one is, in the

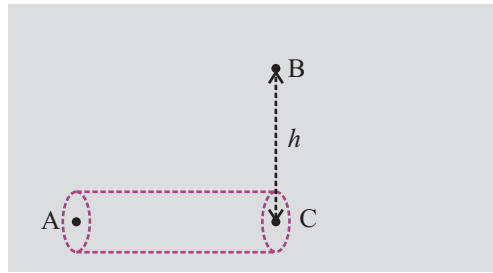


Figure 7-3: Showing points A and B, not in the same vertical line in a liquid at rest, with a vertical separation h between the two (compare with fig. 7-2; C is a point vertically below B at a depth h , so that the line AC is horizontal; a narrow cylindrical portion of the liquid is shown, where the axis of the cylinder is along AC, and its end planes are around A and C, with area $\delta S \rightarrow 0$ each; the equilibrium of this portion of the liquid implies that the pressures at A and C are equal, and hence the pressures at A and B are related by eq. (7-2).

context of the fluid as a whole, an internal force, being made up of a large number of forces, acting through small elements of area in C, and constituting the states of stress at various points of C.

If p be the pressure at any such point, say P, of C and δS a small element of area around P, then the internal force on A exerted by B through this element of area is $p\delta S$ acting along the normal to δS , being directed into the region A. The pressure changes little if one moves slightly away from P to one side, say, to the side occupied by the portion A of the liquid. This pressure, being a measure of the state of stress in the liquid, depends on the state of *strain* which, in turn, is determined solely by the shape and size of the portion A under consideration.

Now imagine the portion B of the liquid to be replaced with some other body, say D, having the *same* interface C with A (fig. 7-4(B)), the entire system being held in equilibrium, if necessary, by forces applied on D. Then the state of strain at every point in A remains the same, and hence the pressure at a point within A close to P continues to be p . Moreover, since the equilibrium of A is not disturbed, all the forces holding it in equilibrium also must remain the same. This means that forces must continue to be exerted by D on A through every small surface element in C as before. In particular, the force through the element δS around P must still be $p\delta S$ along the normal to the area

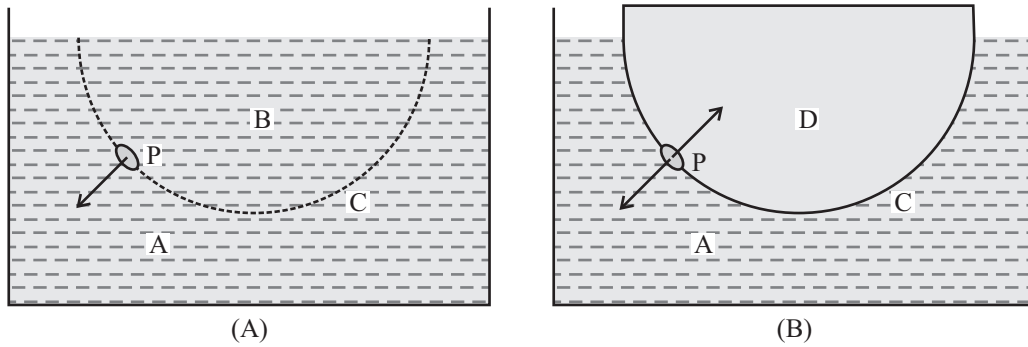


Figure 7-4: Illustrating thrust in a fluid; in (A), C is the imagined interface between two portions A and B of a fluid at rest while in (B) the interface between the body D and the fluid in region A is the same as in (A); the state of stress in region A remains the same in the two situations; the thrust on D exerted by A is then obtained by vector summation of the small contributions of the form $p\delta S$, all directed normally toward the region occupied by D.

δS .

In other words, *regardless of which body is in contact with A through the interface C*, the pressure at a point close to any point P on the interface C equals the pressure (p) at P in the situation depicted in fig. 7-4(A) (i.e., the one where the region B on the other side of the interface C is occupied by the same liquid as in region A), and the force exerted by the body D on the liquid in A through the small surface element δS around P is also given by the same expression (i.e., $p\delta S$) as in the situation in fig. 7-4(A). The latter also gives the force exerted by the liquid in A on the body D through the surface element δS , acting along the normal to the interface at P, but now directed *into* the region occupied by D. The vector sum of all these forces, for all the various area elements δS spanning the interface C, is termed the *thrust* of the liquid in the region A on the body D.

All these considerations apply to any fluid in equilibrium in contact with any other body, say, with a wall of the vessel containing the fluid. The thrust of the fluid on the surface in contact is the total force that the fluid exerts on it, determined by its state of stress at the points adjoining the surface.

Defining in this manner the thrust of a fluid on a wall of the containing vessel as the vector sum of forces of the form $p\delta S$ exerted on small area elements of the wall, one can at the same time obtain the force exerted *by* the walls *on* the fluid, the latter being

simply equal and opposite to the thrust. It is this force on the fluid in the region A in fig. 7-4 that, along with the external forces acting on the fluid particles, and the force exerted by the fluid in region B (or by the body D, as the case may be.), keeps the fluid in the region A in equilibrium.

Fig. 7-5 depicts the cross-section of a vessel filled up to height H with an incompressible liquid of density ρ , where the dimensions of the vessel are as shown, the cross-section by a plane perpendicular to the plane of the figure being a rectangle of width b (which means that the base of the vessel is rectangular with length l and width b). The concepts of pressure and thrust developed above can be used to calculate the thrust on the slanting wall AB of the vessel as well as on the wall CD and on the base BD, as you will see in the next problem.

Problem 7-2

Consider an incompressible liquid in a vessel with a slanting wall AB, the height of the liquid level above the horizontal base BD being H (see fig. 7-5). The section of the vessel perpendicular to the one shown in the figure is rectangular, with width b . If the inclination of the slanting wall to the vertical be θ , find the thrust on the walls AB and CD of the vessel, and on the base BD. Why is the thrust on BD larger than the weight of the liquid in the vessel?

Answer to Problem 7-2

SOLUTION: Refer to fig. 7-5. Imagine a narrow strip (not shown in the figure) of width δy on the slanting wall AB, at a height h above the base, the area of the strip being $b\delta y$. The thrust on this strip is $(p_0 - h\rho g)b\delta y$ in a direction normal to the wall, where $p_0 = H\rho g$ is the pressure at any point on the base of the vessel (assuming for the sake of simplicity that the pressure at the top surface is zero). Since the slanting height y of the strip above the base is related to h as $y = h \sec \theta$, the total thrust is $T_1 = (p_0 b \sec \theta H - \rho g b \sec \theta \int_0^H h dh) = (p_0 b \sec \theta H - \frac{1}{2} \rho g b \sec \theta H^2) = \frac{1}{2} \rho g b \sec \theta H^2$, in a direction normal to the wall. The horizontal component of the thrust is $T_{1H} = \frac{1}{2} H \rho g A$, where $A = Hb$ is the projected area, parallel to the vertical wall CD, of the liquid surface in contact with the slanting wall, i.e., in other words, the area of the liquid surface in contact with the wall CD. It also equals the magnitude of the thrust on CD in a direction normal to it, i.e., along a horizontal direction (check this out). The pressure at all points on the base BD being $p_0 = H\rho g$ the thrust is

$T_2 = H\rho gbl = H\rho gA'$, where $A' = bl$ is the area of the base BD.

Consider now the forces on the mass of liquid contained in the vessel. The wall AB exerts a force on the liquid equal and opposite to the thrust of the liquid on it. This has a horizontal component T_{1H} , pointing from B to D, and a component $T_{1V} = T_1 \sin\theta = \frac{1}{2}H\rho gA \tan\theta$ in the vertically downward direction (i.e., pointing from C to D). The force exerted by the wall CD is T_{1H} along the horizontal direction pointing from D to B, and cancels the horizontal component of the reaction force exerted by the wall AB (check this out). The base of the vessel exerts a reaction force $H\rho gA'$ in the vertically upward direction. Finally, the weight W of the liquid acts vertically downwards. The equilibrium of the liquid contained in the vessel is ensured by the fact that $W + T_{1V} = T_2$ (check this out from the geometry of the vessel), which explains why $T_2 > W$.

NOTE: Check how the results are modified if the pressure on the top surface is, say, P_0 instead of zero (answer: all thrusts are increased by P_0 times the area of the surface in contact).

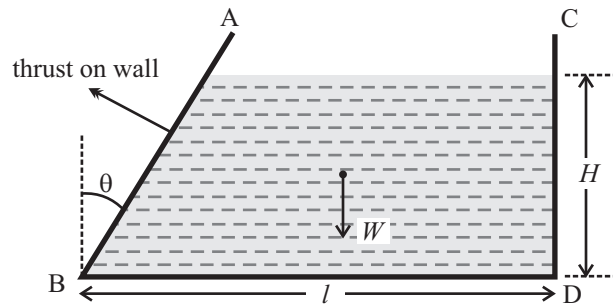


Figure 7-5: Liquid at rest in a vessel with a slanting wall AB, inclined at an angle θ to the vertical; the section by a plane perpendicular to the plane of the figure is a rectangle of width b (thus, the base of the vessel is rectangular in shape with length l and width b); the thrust of the liquid on the slanting wall is along the double-headed arrow; the horizontal component of the force of reaction exerted by the slanting wall on the liquid balances the force of reaction exerted by the vertical wall CD; the vertical component of the force exerted by AB, together with the weight of the liquid, is balanced by the vertical component of the force of reaction of the base on the liquid.

7.2.4 Atmospheric pressure

The atmosphere is made up of a mixture of gases, and can be looked upon as a layer of fluid surrounding the earth, held by the force of gravity. It is, of course, not an incompressible fluid, but a pressure can be defined everywhere in it as indicated in sec. 7.2.1, though the variation of pressure is no longer given by the expression (7-2). A factor con-

tributing to the variation of pressure from one point to another in the atmosphere is the constant flow occurring everywhere in the form of air currents. Still, one can measure the pressure at any point in the atmosphere and, ignoring the variations in pressure from time to time, arrive at an average atmospheric pressure at that point.

One then observes that this average pressure decreases with height as one moves in the vertical direction, in a manner more complicated than that expressed by (7-2). It is convenient to define a *standard atmospheric pressure* at sea level by ignoring certain small variations with distance along the horizontal direction, as also the fluctuations with time. The value of this standard atmospheric pressure is taken to be 1.0133×10^5 Pa. A commonly used unit of pressure, the *bar*, is based on this figure, its value being 10^5 Pa. The standard atmospheric pressure is also sometimes expressed as '760 mm of mercury', or 760 *torr*, one torr being equal to '1 mm of mercury'. This means that if a long evacuated tube be dipped in a vessel containing mercury, and the tube is held vertically with its closed end upward, then the atmospheric pressure will cause a mercury column of height 760 mm to rise in the tube and will hold it in equilibrium (figure7-6).

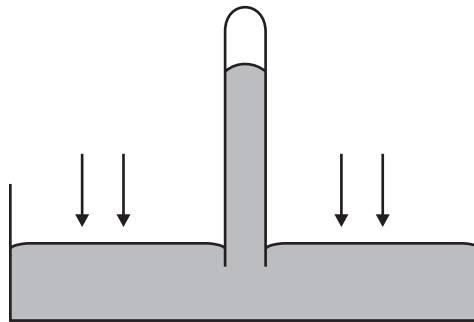


Figure 7-6: Rising of mercury column in an evacuated tube due to atmospheric pressure (schematic); the height of the column can be used as a quantitative measure of the pressure.

Referring to such a column of mercury in an evacuated tube, the pressure at the top of the column has to be zero since the tube was evacuated to start with and so there is no fluid or any other body on top of the column to exert a force on it, and hence, by (7-2) the pressure at the bottom must be $h\rho g$ where h is the column height and ρ stands for

the density of mercury. This must be counterbalanced by the upward thrust per unit area on the column exerted by the atmosphere through the mercury in the vessel, i.e. to the thrust resulting from the atmospheric pressure. Another way of looking at it is that the upward thrust has to be pS where p stands for the atmospheric pressure and S for the area of cross-section of the tube, which must balance the weight of the mercury column in the tube, i.e., $hS\rho g$, giving, once again $p = h\rho g$. Using the known value of ρ and the mean value of the acceleration due to gravity on the surface of the earth, and taking $h = 760$ mm, one arrives at the above value of standard atmospheric pressure.

It is worthwhile to look into the *origin* of pressure in a fluid. As we have seen, pressure is related to the normal force exerted by one portion of the fluid on a contiguous portion through the common surface of separation between the two. Looking at the mechanism underlying the generation of such a force, we can identify two factors which, though related intimately with each other, can, to some extent, be looked upon as distinct causes - (a) the force of interaction between the molecules of the fluid belonging to the two portions of the fluid and located close to the common surface of separation, and (b) the transfer of momentum (see sec. 7.5.3) due to molecules belonging to one portion entering into the other in the course of their random thermal motion.

While the second factor is predominant in the case of a dilute gas, the first one is more relevant in the case of liquids. In the case of a gas at a high pressure, both the factors are to be taken into consideration in accounting for the pressure. In general, the stress force in a fluid possesses a component parallel to the surface of separation and another directed normally to the surface. While the pressure at any point in a fluid is explained in terms of the normal component, the tangential component accounts for the *viscosity* of the fluid (see sec. 7.5.7).

7.2.5 Buoyancy: Archimedes' Principle

Let us refer back to figure 7-4(A), (B), where a part of a fluid at rest, occupying region B in fig. 7-4 (A), has been displaced by a body D in fig. 7-4(B). The latter may or may not be fully immersed in the fluid, i.e., only part of it may be effective in displacing the fluid

in region B, as shown, for instance, in fig. 7-4(B).

A commonly occurring situation is one where the only external force acting on the fluid particles within a region is the force of gravity. In this case the total external force on the fluid in region B in fig. 7-4(A) is the weight of the fluid in this region, which acts through the center of gravity of this portion of the fluid in a vertically downward direction. For the fluid to be at rest, this must be balanced by the internal force exerted on this portion of the fluid by the portion in region A through the interface C. As we have seen, this is *also* the thrust exerted by the fluid on the body D in fig. 7-4(B).

In other words, *the thrust of a fluid on a body fully or partly immersed in a fluid at rest is equal to the weight of the displaced fluid*. This thrust, which necessarily acts vertically upward (assuming that the portion B of the fluid would be in equilibrium if it were not displaced by D), is termed the force of *buoyancy*. Note that this force of buoyancy, acts in a direction opposite to the weight of the body and hence cancels the latter partly or fully. The resulting apparent reduction in weight of the body is equal to weight of the fluid displaced by it. This is known as *Archimedes' principle*.

7.2.6 Equilibrium of fully or partly submerged body

7.2.6.1 Condition of equilibrium

Suppose that the weight of a body partly or fully immersed in a fluid at rest is W while the weight of the fluid displaced by it, which is the force of buoyancy tending to push the body upward, is W' . We assume, for the time being, that the two forces act along the same vertical line and consider one by one the following three possible situations:

(a) $W > W'$, the downward pull on the body due to gravity exceeds the upward push due to buoyancy. The body then cannot be in equilibrium unless an additional external force directed upward is made to act on the body so as to hold it at rest in a partly or wholly submerged position, or else the weight of the displaced fluid (W') is caused to increase. For instance, if the body is partly submerged to start with (fig. 7-7(A)), it may sink in the fluid so as to be submerged to a greater extent (fig. 7-7(B)) as a result of

which W' increases till it becomes equal to W (see below). If the equality of W' and W is not brought about even when the body is fully submerged (fig. 7-7(C)), then it may go on sinking till it presses against some other body like, for instance, the bottom of the containing vessel (fig. 7-7(D)).

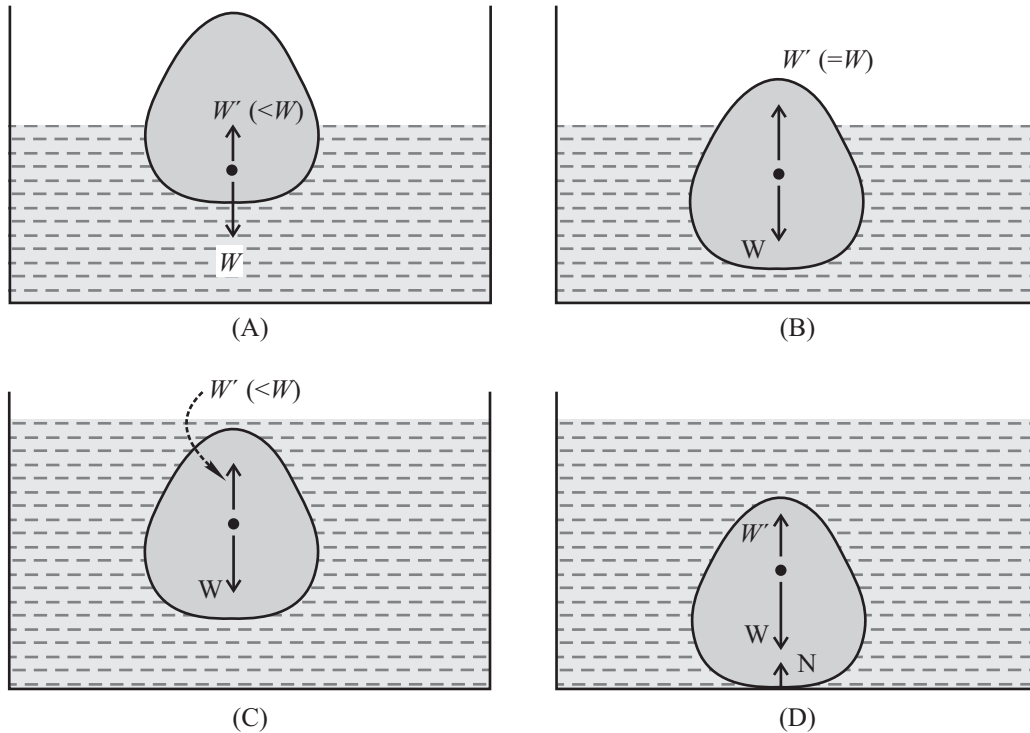


Figure 7-7: Condition of floatation; in (A), the weight W of the body is greater than the buoyancy force W' , as a result of which equilibrium is not possible and the body sinks further and W' increases till it becomes equal to W as in (B); in (C), the equality between W and W' is not brought about even in a fully submerged condition and the body sinks further so as to press against the bottom of the vessel, as in (D).

(b) $W' > W$, the force of buoyancy exceeds the weight of the body. In this case too the body cannot rest at equilibrium and, unless made to do so with the help of an additional downward push applied externally, it has to rise up so that a smaller volume of it remains submerged (fig. 7-8(A), (B)). This corresponds to a smaller volume of fluid displaced, as a result of which, with an appropriate portion of the volume of the body remaining submerged, the equality $W' = W$ may be brought about (see below).

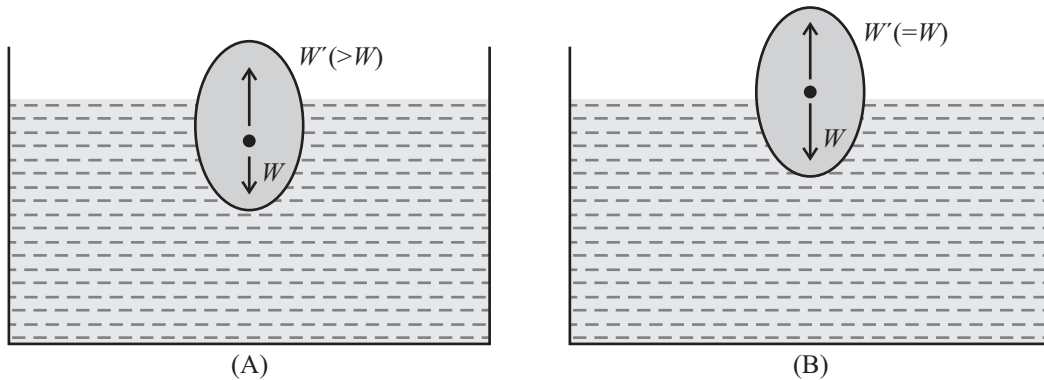


Figure 7-8: Condition of floatation; in (A), the weight W of the body is less than the buoyancy force W' , as a result of which equilibrium is not possible and the body floats up and W' decreases till it becomes equal to W as in (B).

(c) $W' = W$, the force of buoyancy exactly balances the weight of the body. As indicated above, such an equality may be brought about with an appropriate portion of the volume of the body submerged in the fluid (fig. 7-7(B), 7-8(B)). When this happens, the net vertical force on the body is zero. Thus a *necessary* condition for floatation of a body in a fluid is that *the weight of the fluid displaced must be equal to the weight of the body*.

However, the condition $W = W'$, *by itself*, may not be sufficient to ensure equilibrium, since the force of buoyancy resulting from the thrust on the body exerted by the surrounding fluid, along with the weight of the body, may generate a *couple*, tending to *rotate* the body in the fluid. Such a couple arises because the orientation of the body in the fluid may be such that the line of action of the upward force of buoyancy may not coincide with that of the downward force of gravity constituting the weight of the body. While the latter acts through the center of gravity of the body, the former acts through what is termed the *center of buoyancy*, the latter being the center of gravity of the fluid displaced.

Figure 7-9 depicts two situations where the locations of the center of buoyancy and center of gravity, for a body fully submerged in a fluid, are such that the forces W and W' act along two different vertical lines. Hence, even when they are of the same magnitude, they form a couple tending to rotate the body instead of keeping it in equilibrium. In fig. 7-9(A) depicting a body fully immersed in a fluid (such as a balloon floating in air;

the case of a partly immersed body will be considered separately), the couple tends to bring the center of buoyancy and the center of mass in the same vertical line while in fig. 7-9(B), it tends to cause a greater separation between the two along the horizontal direction.

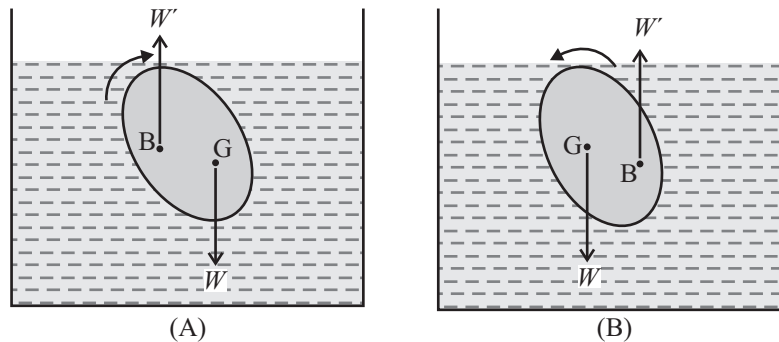


Figure 7-9: Torque generated in a body fully immersed in a fluid; the couple in (A) tends to bring the center of buoyancy B and the center of gravity G in the same vertical line while that in (B) tends to increase the separation between the two along the horizontal direction.

Evidently, then, the following *two* conditions are to be fulfilled in order that the body under consideration may rest equilibrium, when fully submerged, in a fluid (the case of a partly submerged body will be considered below): (C1) the weight of the fluid displaced has to be equal to the weight of the body, and (C2) the center of gravity of the body and the center of buoyancy (denoted by G and B respectively in fig. 7-9(A), (B)) have to lie in the same vertical line (see fig. 7-10(A), (B)) so that the line of action of the force of gravity and that of buoyancy acting on the body may be the same.

7.2.6.2 Stability of equilibrium

Such an equilibrium, however, may or may not be *stable*. In order to explain what this means, I refer to fig. 7-10(A), (B), again showing a body fully immersed in a fluid at rest, where both the above two conditions are fulfilled and the body under consideration rests in equilibrium . In fig. 7-10(A), the center of gravity G lies *below* the center of buoyancy B while, in fig. 7-10(B), G is seen to be *above* B. Suppose now that the body is slightly tilted *away* from its position of equilibrium so as to correspond to the positions

shown in fig. 7-9(A) and (B) respectively. As indicated above, in the former situation the resulting torque on the body tends to restore it to the position of equilibrium while in the latter situation, the body tends to tilt further away from equilibrium.

In other words, the *stability* of equilibrium refers to what happens when the body is slightly tilted away from the equilibrium configuration: in fig. 7-10(A) it tends to regain the equilibrium position when tilted away from the latter, as a result of which the equilibrium is a *stable* one, while, in fig. 7-10(B) it gets tilted further away from equilibrium, as a result of which the equilibrium is *unstable*. It is to be mentioned, however, that, in considering the conditions of equilibrium and of the stability of that equilibrium, we assume that the fluid continues to be at rest, i.e., we ignore the effects of the motion of the fluid relative to the body.

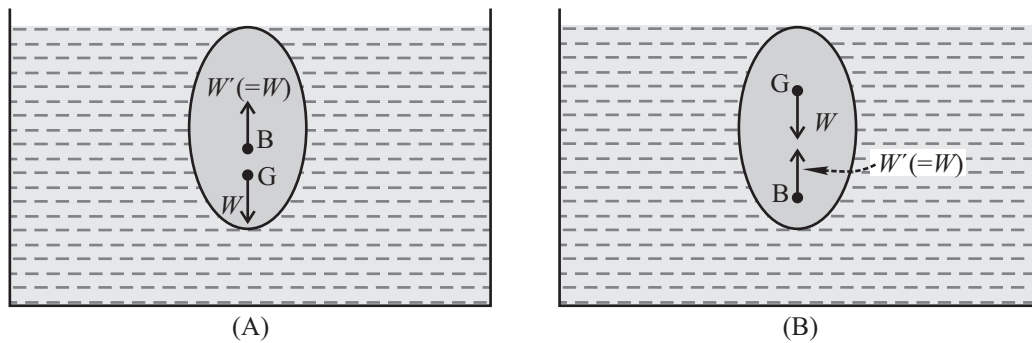


Figure 7-10: Body in equilibrium while fully immersed in a fluid where the two conditions of floatation are satisfied; (A) stable equilibrium and (B) unstable equilibrium.

The case of a body *partly* immersed in a fluid differs from that of a fully immersed one (shown in figures 7-9, 7-10), and requires further considerations. In the latter case, the location of the center of buoyancy relative to the body remains unchanged when the latter undergoes a tilt (reason out why) while, in the former case, the center of buoyancy changes position relative to the body when a tilt occurs (the case of a spherical body is an exception). In order to see whether the equilibrium of such a floating body is stable or not, one has to refer to what is known as its *metacenter*. We consider, for the sake of simplicity, a symmetrical body floating in equilibrium in a liquid (more precisely,

we assume that the body is symmetric about a vertical plane containing the center of gravity, and that the initial tilt occurs about a horizontal axis contained in this plane), with its center of gravity G and center of buoyancy B lying on a vertical line in the plane of symmetry (we refer to this line as the *axis* in the present context). When the body gets tilted in the liquid, the center of gravity continues to lie on the tilted axis with its position fixed in the body, but the center of buoyancy shifts away from the axis to a new position B' relative to the body, as shown in fig. 7-11(A), (B). The vertical line through B' intersects the axis BG at the point M , the metacenter.

One can work out the position of the metacenter for simple geometries (see problem 7-4 for an example; the metacenter can also be determined experimentally), when it is found that it is a point *fixed in the body*, independent of the tilt, with a first order correction proportional to the tilt angle.

If the metacenter M lies above the center of gravity, as shown in fig. 7-11, then the torque on the body, resulting from the force of buoyancy and the weight in the tilted position, tends to restore the body to the upright position which is therefore one of stable equilibrium. On the other hand, if the metacenter lies lower (fig. 7-11), then the body tilts further away, implying that the original upright position is one of unstable equilibrium. Assuming that the equilibrium is a stable one and that the body (assumed to be symmetric about the axial plane referred to above) is released from rest with a small initial tilt, the subsequent motion can be seen to be one of an angular oscillation about the center of gravity *along with* a linear oscillation of the latter in the vertical direction.

1. All the results of the present section are conditional to the assumption that the fluid around the body under consideration remains at rest. In reality, the oscillatory motions of the body disturb the equilibrium of the fluid as well, which may additionally possess a flowing motion of its own.
2. For sufficiently small tilt angles, the vertical and the angular oscillations are independent of each other while, in the next order of approximation, they are seen to be coupled to each other, i.e., one cannot occur independently of the other.

3. In the case of stable equilibrium, where the metacenter M lies above G , the angular oscillations for a small initial tilt are in the nature of simple harmonic one, with the angular equation of motion of the form of eq. (4-28), and the time period of oscillation is seen to work out to (see, for instance, problem 7-4)

$$T = 2\pi\sqrt{\frac{I_G}{mgl}}, \quad (7-3a)$$

where l stands for the distance MG , referred to as the *metacentric height*, m for the mass of the floating body, and I_G for the moment of inertia of the body about the axis of oscillation, which passes through the center of gravity. If the radius of gyration K of the body about the metacenter M be such that $l \ll K$ (this condition is satisfied, for instance, in the case of a ship) then one can, in an approximate sense, replace I_G in the above expression with I , the moment of inertia about a displaced axis passing through the metacenter M , parallel to the axis of rotation through G . In other words, the time period of oscillation is effectively the same as that of angular oscillations of an equivalent compound pendulum suspended from this displaced axis, and one has

$$T \approx 2\pi\sqrt{\frac{I}{mgl}}. \quad (7-3b)$$

In this approximation, the time period is inversely proportional to the square root of the metacentric height. Thus, if the metacentric height is made to increase, the oscillations become more stable (i.e., for a given tilt angle, the restoring couple becomes larger), with an attendant decrease in the time period. This assumes importance, for instance, in the design of a ship where it is desirable to increase the degree of stability and, at the same time, to ensure that the time period does not become too small (since this would cause discomfort to passengers).

4. For sufficiently small tilt angles, the vertical oscillations are also simple harmonic in nature, whose time period is the same as the time period of vertical oscillations in the upright position.

Notice that in fig. 7-11(A), the center of gravity G is shown to lie *above* the center of buoyancy B in the upright position, and still the equilibrium is a stable one owing to the fact that the metacenter is above the center of gravity (recall that, for a fully immersed

body with G lying above B , the equilibrium is unstable). On the other hand, for a floating body with G lying above B , but with the metacenter M lying below G , the equilibrium is unstable, analogous to a fully immersed body. Finally, with G lying below B , the metacenter is always above G , and the equilibrium is stable, again analogous to a fully immersed body. Indeed, for a fully immersed body, the metacenter and the center of buoyancy are identical points.

In summary, the conditions (C1) and (C2) stated above are necessary for the equilibrium of a body in a fluid while, for the equilibrium to be stable, it is further required that the center of gravity should lie below the metacenter (which coincides with the center of buoyancy for a fully immersed body).

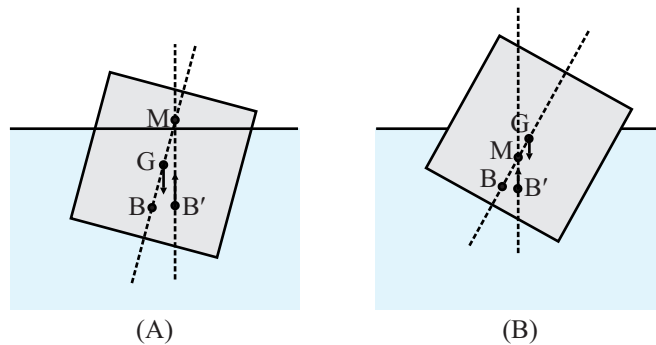


Figure 7-11: Body, partly immersed in a liquid, illustrating (A) stable (B) unstable configuration (schematic); in (A) the metacenter M is located above the center of gravity G , in which case it performs stable angular oscillations about G when tilted slightly away from the equilibrium position (additionally, there occurs a vertical oscillation of the center of gravity); the distance MG is the metacentric height; in (B) the metacenter is below the center of gravity and the angular motion is an unstable one; in both the cases, a tilted position of the body is shown, along with the directions of the force of gravity and that of buoyancy acting on the body through G and B' respectively, where B' represents the displaced center of buoyancy; M is the point of intersection of the vertical line through B' with the tilted axis GB , B being the position, with respect to the body, of the center of buoyancy in the upright position of equilibrium.

The above statements, however, do not exhaustively answer the question of stability of floating bodies because a complete answer involves a number of additional considerations. First of all, even though a body is unstable against a small tilt, there may be restoring mechanisms at larger tilts. On the other hand, stability against small tilts does not preclude instability at larger ones. The stability of a ship, for instance,

depends in a complex way on the mass distribution in the ship. The ship may have various different *modes* of tilting motion like rolling and pitching, and the stability of the various different modes (against small and large tilts) may depend on a number of competing factors. Finally, the motion of the fluid itself exerts a great influence in determining the stability since a body moving through a fluid experiences a number of forces on account of the relative motion between the two. The stability of a ship in the sea or of an aircraft in the atmosphere are, therefore, problems of a complex nature.

Problem 7-3

A wooden cylinder of length $l_1 = 0.5$ m and a metal cylinder of length $l_2 = 0.02$ m, each with a radius $R = 0.1$ m, are tied together to form a composite cylindrical body with a common axis, which floats in water with the wooden cylinder on top of the metal one, and the cylinder axis vertical (see fig. 7-12). If the densities of water, wood, and metal be $\rho_0 = 1.0 \times 10^3$ kg·m⁻³, $\rho_1 = 0.6 \times 10^3$ kg·m⁻³, and $\rho_2 = 8.0 \times 10^3$ kg·m⁻³ respectively, what length of the composite body is submerged in water? Where is the center of buoyancy of the composite floating body located with reference to its center of gravity? Calculate the force exerted on the wooden cylinder by the metal cylinder by means of the joining device (say, a piece of thread).

Answer to Problem 7-3

Refer to fig. 7-12. If the required length of the composite cylinder submerged in water be l (see fig. 7-12) then, according to the condition of floatation (which requires that the weight of the composite body must be equal to the weight of the water displaced), $\rho_0 l = \rho_1 l_1 + \rho_2 l_2$ (reason this out), which gives $l = 0.46$ m.

The center of gravity G of the composite cylinder is at a distance d_1 from the top where $d_1 = \frac{l_1^2 \rho_1 + l_2(l_1 + \frac{l_2}{2})\rho_2}{l_1 \rho_1 + l_2 \rho_2}$ (reason this out), i.e., $d_1 = 0.34$ m. The center of buoyancy G' is located at the point where the center of gravity of the displaced water would be, whose distance from the top of the composite cylinder is $d_2 = (l_1 + l_2 - \frac{l}{2})$ (reason this out), i.e., $d_2 = 0.29$ m. In other words, the center of buoyancy is at a distance 0.05 m vertically above the center of gravity.

The buoyancy force on the metal cylinder is $F_B = \pi R^2 l_2 \rho_0$, acting vertically upward, while its weight is $W = \pi R^2 l_2 \rho_2$ acting vertically downward. Hence, for the equilibrium of the metal cylinder,

a force $F = W - F_B$ has to act on it in the vertically upward direction due to its coupling with the wooden cylinder. This implies that an equal and opposite force is exerted on the wooden cylinder. Making use of appropriate values, this works out to $F = 4.4 \times 10^{-3}$ N in the vertically downward direction.

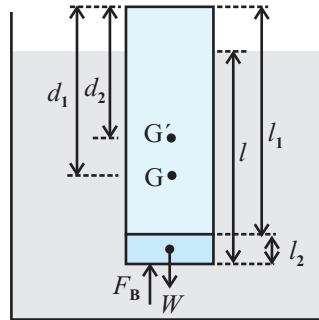


Figure 7-12: A composite cylindrical body made of two cylinders of length l_1 and l_2 and radius R each, floating in a liquid of density ρ_0 (refer to problem 7-3); the densities of the materials of the two bodies are ρ_1 and ρ_2 ; the axis of the composite cylinder is vertical with the first cylinder on top of the second; the condition of floatation can be invoked to determine the length l of the submerged portion of the composite cylinder; the center of buoyancy G' is located at the point at which the center of mass of the displaced liquid would have been.

Problem 7-4

A homogeneous cubical body of edge length a , made of a material of density ρ ($< \rho_0$, the density of water), floats in water and is released from rest when tilted slightly away from the upright position of equilibrium, accompanied by a small displacement of the center of gravity in the vertical direction. Find the position of the center of buoyancy (in the upright position and also in an inclined position with a small tilt θ), and of the metacenter. Show that the equilibrium is stable and find expressions for the metacentric height and the time period of angular oscillations. Determine the time period of the vertical oscillations of the center of gravity.

Answer to Problem 7-4

HINT: Fig. 7-13 shows a cross-section of the the cubical body but, for the sake of convenience of presentation, in a rotated co-ordinate system where the water level L is shown inclined and ABCD depicts the cross-section of the body in the tilted position (hence the upright position of the cube, not shown in the figure, would appear inclined toward the left by an equal and opposite

angle of tilt). The height h_0 of the water level above the center of gravity in the upright position in equilibrium (not shown in the figure) is $h_0 = a(\frac{\rho}{\rho_0} - \frac{1}{2})$. In the displaced position (which, in general, involves a tilt and also a vertical displacement of the center of gravity), the heights of the water level on the two sides of the block (points P and Q on AD and BC) above the displaced center of gravity (which, however, is fixed in the body) are, say, $h_0 + x$ and $h_0 + x - y$, where $y = a \tan \theta$, θ being the angle of tilt. A co-ordinate system fixed in the body is shown, with the origin O coinciding with the center of gravity G, and axes OX, OY, where the latter is shown vertical, with the water level L making an angle θ with OX. The co-ordinates of a number of points relevant for the problem are shown.

The volume of water displaced (cross-section PQCD) is $a^2(\frac{a}{2} + h_0) + a^2x - \frac{a^2}{2}y$, and the net upward force is $a^2(x - \frac{a}{2}u)\rho_0g$, where $u = \tan \theta = \frac{y}{a}$ is a small parameter ($\approx \theta$) specifying the tilt, and g stands for the acceleration due to gravity.

Considering the part of the cube within the triangular cross-section PQE, the mass of water in it is $\frac{1}{2}\rho_0a^2y$, and its center of mass is at $(0, -\frac{a}{4} + \frac{h_0+x-y}{2})$. Similarly, considering the portion with rectangular cross-section PECD, the mass of water is $\rho_0a^2(\frac{a}{2} + h_0 + x - y)$, and its center of mass is at $(0, -\frac{a}{4} + \frac{h_0+x-y}{2})$. Thus the total mass of water displaced is $m = a^2\rho_0(\frac{a}{2} + h_0 + x - \frac{y}{2})$, and the location of the center of buoyancy (B') in the displaced position of the body is at (p, q) (say), where, up to the first order of smallness, one finds $p \approx \frac{\rho_0 y}{12\rho}$ (making use of the expression for h_0), and $q \approx q_0(1 - \frac{a(x - \frac{y}{2})}{\frac{a^2}{4} - h^2})$, $q_0(= \frac{a}{2}(\frac{\rho}{\rho_0} - 1))$ being the y-co-ordinate of the center of buoyancy B in the undisplaced position (thus the center of buoyancy in the upright position of equilibrium lies below the center of gravity; the corresponding x- co-ordinate is $p_0 = 0$).

The net upward force due to buoyancy in the displaced position is given by $w \equiv mg - Mg \approx \frac{a^2}{2}g\rho_0(2x - y)$ ($M = a^3\rho$ is the mass of the body), and the vertical displacement of the center of gravity of the body works out to (approx) $\zeta = x - \frac{y}{2}$ (check all these statements out).

The metacentric height (distance from G to M) now works out to

$$l = \frac{p}{u} + q_0 = \frac{a\rho_0}{12\rho} - \frac{a}{2}(1 - \frac{\rho}{\rho_0}) > 0, \quad (7-4)$$

where terms of the first degree of smallness are ignored (check this expression out). This shows that the metacenter lies above the center of gravity (i.e., the equilibrium is a stable one) and is a

fixed point in the body when first degree terms are ignored. The moment of the couple tending to increase the tilt is $-mgl \sin \theta \approx -mgl\theta$. The angular equation of motion describing the rotation about the center of mass of the body is thus

$$I_G \frac{d^2 \theta}{dt^2} = -mgl\theta, \quad (7-5a)$$

where I_G stands for the moment of inertia about the center of mass. This implies angular oscillations of the simple harmonic type, with a time period

$$T = 2\pi \sqrt{\frac{I_G}{mgl}}. \quad (7-5b)$$

The equation of motion describing the vertical oscillation of the center of mass is seen to be

$$m \frac{d^2 \zeta}{dt^2} = -a^2 g \rho_0 \zeta, \quad (7-6a)$$

corresponding to which the time period is

$$T' = 2\pi \sqrt{\frac{a\rho}{g\rho_0}}, \quad (7-6b)$$

which is independent of the angular motion in the approximation employed.

Ignoring terms of the second and higher degrees in the angle of tilt, the expression for the metacentric height given in (7-4) can be generalized to

$$l = \frac{J}{V} - l_{GB}, \quad (7-7a)$$

where V stands for the volume of the liquid displaced, l_{GB} for the distance from the center of gravity to the center of buoyancy in the upright position, and J for the integral

$$J = \int_S x^2 dx dz'. \quad (7-7b)$$

In this integral, $dx dz'$ represents an infinitesimal element of area around the point with co-ordinates (x, z') in the horizontal cross-section (denoted by the symbol S) of the floating body taken at the level of the surrounding liquid, where the z' -axis is chosen parallel to the axis OZ (passing through the origin O , i.e., the center of gravity

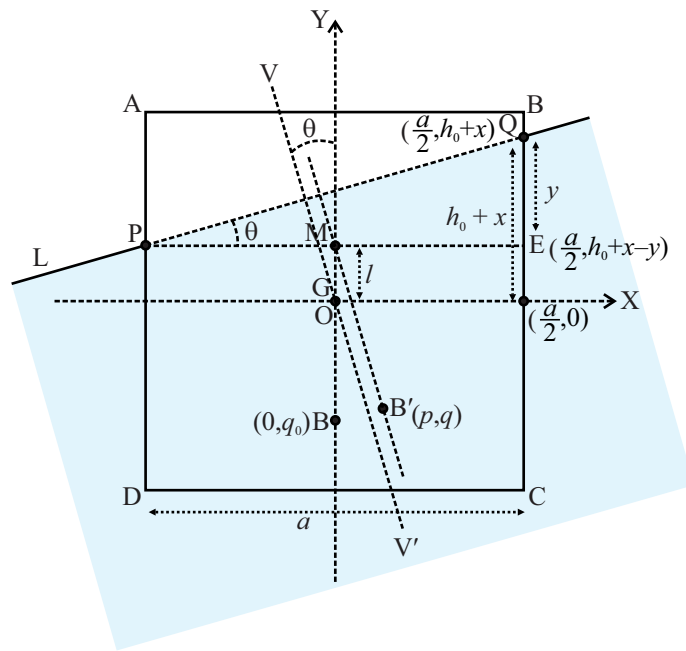


Figure 7-13: Showing a cubical body, of edge length a , and made of a material of density $\rho (< \rho_0)$, floating in water (density ρ_0), where a *rotated* view is presented for the sake of convenience, in which ABCD represents the body in a *tilted* position, and L represents the water level relative to this tilted position, the angle of tilt being θ (refer to problem 7-4). The line VV' represents the vertical line, which now appears tilted, in the upright position of the body (not shown in this figure); OX, OY are co-ordinate lines fixed in the body, with respect to which B represents the position of the center of buoyancy in the upright position, while the center of gravity G coincides with O; B' represents the shifted center of buoyancy relative to this body-fixed co-ordinate system; the line through B', drawn parallel to VV' intersects the line BG in M, the metacenter, which lies *above* G in the present instance, implying that the upright position of equilibrium is a stable one; distances shown are not to scale.

of the body; not shown in fig. 7-13) about which the angular oscillations take place, lying in the same vertical plane as the latter.

In writing the formula (7-7a) we have again assumed that the floating body is symmetric about the y - z plane, i.e., about the vertical plane mentioned above, which contains the axis of oscillation and the unperturbed center of buoyancy.

7.2.7 Pascal's law: transmission of pressure

Notice that eq. (7-2) involves only the *difference* in pressure at any two points in an incompressible liquid, and not the pressure in itself. If the pressure is known at any one point in the liquid, then its value at any other point gets determined from (7-2). One

way to fix the pressure at any point in the liquid, and not just the pressure difference, is through the force per unit area at some *boundary* point in it, where the liquid is in contact with some other body, say, with part of the containing vessel.

Fig. 7-14 depicts a situation where a vessel is completely filled up with an incompressible liquid, and the point A in the liquid is adjacent to a piston of cross-section A , on which an external force of a desired magnitude (say, W) can be applied, the latter acting inward, as shown. Evidently then, if the piston be in equilibrium under this applied force and the thrust on it exerted by the liquid, the pressure at A is related to A and W as

$$p_A = \frac{W}{A}, \quad (7-8)$$

(check this out). Once you know the pressure at A, you can use eq. (7-2) to determine the pressure at any other point, in the liquid. Thus, for instance, the pressure at the point B, in contact with some other part of the containing vessel, would be

$$p_B = p_A + h\rho g = \frac{W}{A} + h\rho g, \quad (7-9)$$

where h stands for the vertical height of A above B. Now suppose that the external force on the piston adjacent to A is increased by w and the liquid is allowed to come to equilibrium with this added force on the piston. This results in an increase in the value of p_A , the pressure at A, by an amount $p = \frac{w}{A}$. Eq. (7-9), written with this increased value of p_A then tells us that p_B also gets increased by the *same* amount, since h and ρ remain unchanged, the latter because the liquid is assumed to be an incompressible one. This added pressure at B would result in an increased thrust on the wall of the containing vessel adjacent to it, acting normally to the wall. Moreover, since the point B has been chosen arbitrarily, the *same* increase in pressure results *everywhere* in the liquid, irrespective of its location. One thus arrives at *Pascal's law* for an incompressible liquid:

In a vessel completely filled with an incompressible liquid in equilibrium, if the pressure at any point in the liquid be increased by any given amount, say, p , then the same increase in pressure is transmitted everywhere in the liquid and results in an increased thrust on any

surface in contact with it, the increment in thrust per unit area being p , acting normally to the surface under consideration.

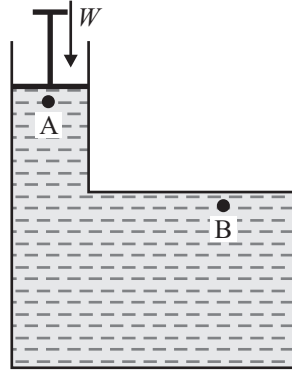


Figure 7-14: Illustrating Pascal's law; an additional pressure generated at A is transmitted undiminished to B; the liquid is assumed to be incompressible.

A condition for this law to hold is that the external forces acting on all the volume elements in the liquid are to remain the same. For instance, in the situation we have considered here, the same gravitational forces act on the volume elements both before and after the increment of force on the piston at A, as a result of which the second term on the right hand side of eq. (7-9) remains the same.

The assumption of an incompressible liquid is an idealized one and implies that an additional pressure generated anywhere in the liquid is transmitted instantaneously, and undiminished, through its volume. In reality, an applied additional pressure propagates through a fluid by means of a *pressure wave* generated in it. Pressure waves of small amplitude will form the subject matter of chapter 9.

The above principle is made use of in numerous appliances in the form of *principle of multiplication of force*, the latter being illustrated in fig. 7-15 which is similar to fig. 7-14, but where the point B is adjacent to a second piston of area A' fitted in the wall of the containing vessel. In this case, the increment in thrust per unit area on this second piston being $\frac{w}{A}$, the total thrust on it will get increased by $\frac{wA'}{A}$. Thus, an increase in force

(applied downward in fig. 7-15) by an amount w on the piston adjacent to A results in an increase in force (acting upward in the situation shown) on the piston adjacent to B by an amount $w \frac{A'}{A}$.

In other words, the applied force w gets multiplied by the ratio $\frac{A'}{A}$, generating the force on the second piston at B, this being the ratio of areas of the two pistons. For instance, if the piston at B has an area 10 times that of the piston at A, and if an additional force of 1 N is applied to the latter, then the incremental force exerted on the former due to transmission of pressure will be 10 N. An appliance of great importance and versatility working on this principle of multiplication of force is the *hydraulic press*.

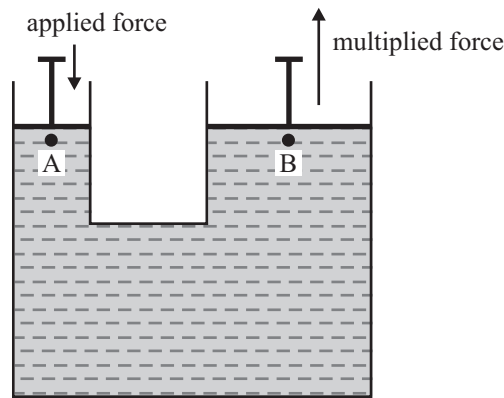


Figure 7-15: Principle of multiplication of force; an additional force applied to the piston adjacent to A results in a multiplied force being exerted on the piston adjacent to B.

With a force w applied to the piston at A in fig. 7-15, the system can be in equilibrium only if a force $w \frac{A'}{A}$ is applied downward on the other piston at B, this being equal and opposite to the increased thrust exerted by the liquid on the latter.

Supposing that the first piston is made to move down by x , the incompressibility of the liquid implies that the second piston will move upward by $x \frac{A}{A'}$ (check this out). In other words, the distance x gets multiplied by the *reciprocal* ratio as compared to the ratio through which the force is multiplied. For instance, in the above numerical example, while the force gets multiplied by the ratio 10, the second piston will move through a distance $\frac{1}{10}$ times that covered by the first one. This means, in turn, that the work done

by the force applied on the first piston is the *same* as the work done *on* the second piston: while the force gets multiplied, *energy remains conserved*.

7.3 Fluids in motion: a few introductory concepts

As in the mechanics of particles and rigid bodies, one usually distinguishes, in the context of mechanics of fluids, between problems in *hydrostatics* and those in *fluid dynamics*. While the above paragraphs refer essentially to fluids at rest, the principles underlying the *flow* of fluids are also of great relevance.

Fluid dynamics in its present day form is the result of a convergence of two streams of investigation. The science of *hydraulics* was developed by practical men in the fields of irrigation, water supply, river management, ballistics, balloon flight and similar other areas of application, while the theoretical principles of *hydrodynamics* were developed by mathematicians and physicists by considering simplified and abstract models of fluid flow. The fruitful exchange between the two streams and their subsequent merger has made possible great advances in applied sciences like those of aerodynamics, meteorology and oceanography, as also in several theoretical areas in mathematics and physics.

A description of the motion of real fluids in all generality is, to all intents and purposes, an impossible task, being one of enormous complexity. A more modest and practical approach is to consider idealized models of fluid flow where a number of features of actual flows are ignored, retaining only a few others for consideration. This is to be done judiciously so that the simplified flow so considered retains some measure of relevance from the practical point of view and features of the flow can be tested by appropriate means.

To start with, one distinguishes between *laminar* and *turbulent* flows, where the former is simpler from the conceptual point of view while turbulence is a more complex feature to be found in almost all actual flows. A laminar flow can be described in terms of the configuration of *stream lines* (see below) and of their motion, while in a turbulent flow the stream lines break up and do not have well defined structure. A laminar flow,

moreover, can be either a *steady* or a *time dependent* one - a classification based on the time dependence of the configuration of the stream lines. Another relevant classification of laminar flows divides these into *rotational* and *irrotational* ones. Considering the two classifications together, one can have several types of laminar flow, of which *steady irrotational* flows are relatively simple to describe. I will now very briefly explain to you the meaning of all these terms, but before that a few more comments are in order so as to make clear what kinds of idealization are involved in the consideration and description of each specific instance of fluid flow.

An actual fluid is characterized by a number of physical properties like compressibility, thermal conductivity, viscosity, and surface tension. In any given context, one or more of these properties may not be of essential relevance in determining the flow characteristics and may be ignored so as to lead to a simplified flow model that, at the same time, is adequately representative of the flow in the given context. For instance, in describing the flow of a liquid, it can be assumed to be *incompressible* in numerous situations of interest, especially when the generation and propagation of sound waves and shock waves are not of any significant relevance. Again, in describing a fluid flow in regions away from solid surfaces one can assume that the fluid is an *ideal* one, i.e., its viscosity is negligible, provided that the typical velocity characterizing the flow is sufficiently small since, in such a situation, one arrives at a reasonably correct description of the flow, at least in a qualitative sense, in spite of the idealization. Finally, if the temperature distribution in the fluid is sufficiently uniform and if the effects of turbulence are not pronounced, one can assume that the thermal conductivity of the fluid is zero and still arrive at meaningful conclusions regarding the flow. The effects of turbulence and a non-zero thermal conductivity lead to *energy dissipation* within the bulk of the fluid, even when the viscosity is small in magnitude.

7.3.1 Stream lines: steady flow

We first define the term *stream line*. A stream line in a fluid at any given instant of time is a line such that the tangent at every point on it gives the direction of velocity of the fluid particle located at that point at that time instant. Fig. 7-16 depicts schematically

a set of stream lines in a flowing fluid, with the direction of velocity shown at three different points (A, B, C) on a stream line.

A flow is said to be of the *laminar* type if, at every instant of time during an interval, the family of stream lines is arranged in a regular pattern so that the velocity $\mathbf{v}(\mathbf{r}, t)$ of a fluid particle at the position \mathbf{r} and at time t is a well defined and regular function of \mathbf{r} and t . In this case, the motion of the fluid near any given point can be locally described as the sliding of successive layers of the fluid over one another.

1. The term 'fluid particle' needs explanation. While the fluid is made up of microscopic constituents like the atoms and molecules, it is not these constituents that one refers to by the term 'fluid particle' here. Instead, the term refers to a small mass of the fluid of *macroscopic* proportions where the mass is *imagined* to be vanishingly small, again in macroscopic terms. Thus, when speaking of the velocity of a fluid particle, one does not mean the velocity of a single molecule but the average velocity of the molecules making up a small mass element of the fluid. Thus, for instance, an element of volume, say, of linear dimension $\sim 10^{-6}$ m in a liquid may be considered to be so small as to be effectively treated like a particle while at the same time containing a very large number of molecules and qualifying as a macroscopic object. The molecules in such a volume element may have their actual velocities spread over a wide range since the molecular velocities involve what is referred to as *thermal* fluctuations, i.e., variations depending on the temperature of the liquid. The average over these variations of the velocities of the individual molecules belonging to a small mass element located at any point in a fluid is then defined as the velocity of a fluid 'particle' at that point.
2. In considering the mechanics of fluids, we look at a fluid as a *continuous medium* while, in reality, the fluid is made up of molecular and atomic constituents. All physical quantities relating to the fluid are interpreted as averages over small but macroscopic volume elements at specified space and time co-ordinates. These macroscopic quantities are related to microscopic ones in the form of appropriate averages within the framework of *kinetic theory*.

Looking at a fluid as a continuous medium, the velocity vectors of the fluid particles lo-

cated at all the various points in a region of space at any given instant of time constitute a *vector field* (refer to sec. 2.13) referred to as the *velocity field* of the fluid, that may or may not be a time dependent one. In the latter case, which is more general, the stream lines change their shape and move through space as the flow continues. For instance, the dotted line in fig. 7-16 shows schematically the changed configuration of the stream line ABC at a later instant.

If, however, the flow be such that the stream lines *remain unaltered* in time, i.e., the velocity field is time-independent, then it is termed a *steady flow*. Thus, if the flow depicted in fig. 7-16 is a steady one then the stream line ABC (or any other stream line for that matter) will remain unaltered in course of time, and the dotted line showing its configuration at a later time will then evidently coincide with ABC itself. Further, the paths followed by the fluid particles in course of time in a steady flow are the stream lines themselves. For instance, the particle located at A in fig. 7-16 at any given instant will follow the path ABC as the flow continues. However, a steady flow can, in general, be *non-uniform* in nature, i.e., fluid particles can be accelerated in the course of their motion and the distribution of stream lines in one region of space may differ from that in some other region.

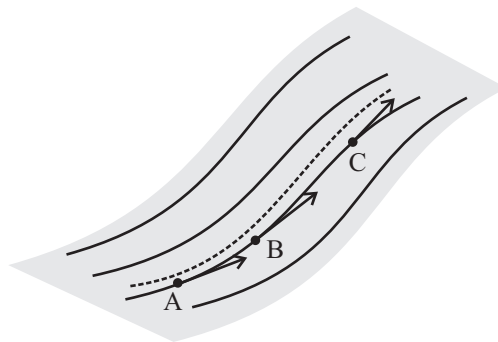


Figure 7-16: Stream lines in a fluid; a set of stream lines is shown; velocities of fluid particles located at various points of a stream line ABC are indicated with arrows; the dotted line indicates possible altered configuration of ABC at a later instant.

7.3.2 From laminar flow to turbulence

Referring to the laminar flow introduced above, a steady flow is necessarily laminar while time dependent laminar flows are more general. For instance, one may refer to a flow where the stream lines undergo a periodic change of configuration, as observed in certain convective flows of fluids.

On the other hand, the stream lines may break up into tiny fragments and the velocities of fluid particles may change in an almost *random* manner with time, making the velocity depend irregularly on position and time. This corresponds to what is known as *turbulent flow* (see sec. 7.5.4 for further considerations). Under certain circumstances, there occurs a *transition* from laminar to turbulent flow in a fluid as certain parameters characterizing it are made to change in appropriate ways (refer to sec. 7.5.9), in which the *viscosity* of the fluid plays a role.

7.3.3 Rotational and irrotational flows

The classification of laminar flows into rotational and irrotational ones is of quite considerable significance.

The curl of the vector field $\mathbf{v}(\mathbf{r}, t)$ at any point (see section 2.14.1 for the definition of the curl of a vector field) in a flowing fluid is referred to as the *vorticity* at that point. If the vorticity at every point throughout the volume occupied by the fluid is zero, then the flow is said to be an irrotational one. On the other hand, if the curl does not vanish identically, the flow is said to possess a rotation. In the case of an irrotational motion, if one considers any arbitrarily chosen fluid element, then that element will be found to possess no rotational motion about its center of mass.

An irrotational flow is relatively easy to analyze and describe compared to a flow with rotation, since, for an irrotational flow, there exists a velocity *potential*, i.e., a scalar field ϕ in terms of which one can express the velocity field as the gradient of ϕ (refer again to sec. 2.14.1 for the definition of the gradient of a scalar field).

In the case of an *ideal* fluid, i.e., a non-viscous one (in reality, a fluid always possesses

viscosity, though in some situations the effects of a non-zero viscosity may be negligible; in addition, an ideal fluid is assumed to have zero conductivity), vorticity cannot be created or destroyed in the course of fluid flow, i.e., if the flow is irrotational to start with, it continues to be irrotational, and likewise, a flow with non-zero rotation remains so. Since irrotational flows are relatively easy to describe, these can be made use of, in approximation schemes, as starting points from which flows with rotation can be constructed for fluids where viscosity effects are small.

7.3.4 Equation of continuity

The flow of a fluid involves transport of mass, momentum, and mechanical energy (made up of kinetic and potential energies) from one region of space to another and from one fluid element (with some small volume) to another. In addition, there occurs, generally speaking, *dissipation of energy* within the fluid, caused by viscosity and thermal conduction. A flow is said to be ideal if the dissipation effects are of negligible relevance.

Generally speaking, electrical and magnetic properties of a fluid may also be involved in a flow in a non-trivial way, as in the case of a plasma. However, we do not consider flows of such generality in this brief introduction.

Assuming that there occurs no chemical transformation within the fluid, the *conservation of mass* is expressed mathematically in the form of what is referred to as the *equation of continuity*. Considering a volume within the fluid with a fixed boundary, the equation of continuity expresses the fact that the rate of increase of mass within this volume must equal the rate at which the fluid, carrying mass, enters into this volume through the boundary surface. In the case of an *incompressible* fluid (i.e., one for which the density is a constant, independent of the pressure), this equation assumes the following form involving the velocity field $\mathbf{v}(\mathbf{r}, t)$ of the fluid

$$\frac{\partial \mathbf{v}}{\partial t} + \text{div } \mathbf{v} = 0. \quad (7-10a)$$

If, moreover, the flow is a steady one, then the equation of continuity assumes a still

more simple form:

$$\text{div } \mathbf{v} = 0. \quad (7-10b)$$

This equation for continuity for the steady flow of an incompressible fluid can be expressed in a particularly simple form in the case of flow through a fixed tube since, then, the volume of liquid flowing through any cross section of the tube must be a constant, independent of where and how that cross section is taken (see problem 7-5 for a simple instance).

In this chapter, unless otherwise stated, various features of fluid flow will be explained with reference to incompressible fluids (i.e., liquids), though the term 'fluid' will be retained while referring to features that hold, in at least a qualitative sense, for compressible fluids as well.

7.3.5 Ideal fluid: equation of motion

Considering a small volume element of a fluid, the rate of change of momentum of this element in the course of its motion can be worked out from the forces acting on it, where the forces can be of different origin. For instance, one commonly occurring force is that of gravity, which can be expressed in terms of the gravitational potential energy of the element under consideration. More generally, the fluid can be subjected to a conservative force field, for which the potential energy per unit mass of the fluid at any given point is, say $\phi(\mathbf{r})$ where the force field is assumed to be time independent.

The volume element of fluid under consideration experiences another force due to its interaction with contiguous elements occurring through its boundary surface. One part of this force acts normally at every point of the boundary surface, being directed into the interior of the volume element, whose value per unit area around any given point is just the *pressure* at that point. On considering all the points on the boundary surface, the resultant force per unit volume due to the pressure distribution in the fluid turns out to be of the form $-\text{grad } p$, i.e., the gradient of the pressure field with a negative sign.

A third component of the force on the volume element under consideration is exerted *tangentially* at every point of the boundary surface, due to the *viscosity* of the fluid, and is responsible for energy *dissipation* in the fluid. For the sake of simplicity, however, we will ignore dissipative effects, noting that there exist flows where such simplification still yields meaningful results, at least in respect of a number of principal flow characteristics.

With this simplification, the total force per unit mass acting on the volume element works out to $-\text{grad } \phi - \frac{\text{grad } p}{\rho}$, where ρ stands for the density at any specified point. This must be equal to the rate of change of momentum per unit mass. In the case of a stationary volume element, this would be given by the expression $\frac{\partial \mathbf{v}}{\partial t}$. In general, however, the element moves in space, and the correct expression for the rate of change of momentum per unit mass is found to workout to $\frac{\partial \mathbf{v}}{\partial t} + \vec{\Omega} \times \mathbf{v} + \frac{1}{2} \text{grad } v^2$, where $\vec{\Omega} = \text{curl } \mathbf{v}$ stands for the vorticity introduced in sec. 7.3.3.

Accordingly, one obtains the following *equation of motion for an ideal fluid*

$$\frac{\partial \mathbf{v}}{\partial t} + \vec{\Omega} \times \mathbf{v} + \frac{1}{2} \text{grad } v^2 = -\frac{\text{grad } p}{\rho} - \text{grad } \phi. \quad (7-11)$$

A simpler form of this equation, with the second term on the right hand side absent, describes the motion of an ideal fluid in the absence of external forces, and is referred to as *Euler's equation of motion*.

Eq. (7-11), though written for the special case of an ideal fluid, is of considerable relevance nevertheless, from which a number of useful conclusions can be derived. For instance, in the case of a fluid at rest, the first two terms on the left hand side reduce to zero and the remaining terms can then be made use of to derive a relation between pressure, density, and the potential due to the external force. In the particular case of an incompressible liquid in a uniform gravitational field this leads to the expression (7-2) for pressure difference between any two points in the liquid derived earlier from more elementary considerations. Pascal's law (refer to section 7.2.7) then follows as a corollary.

Again, for a *steady* flow, the first term on the left hand side of (7-11) drops out, and one can then arrive at *Bernoulli's principle* which we derive below from more elementary considerations (see sec. 7.3.6, where the special case of an incompressible liquid in a homogeneous gravitational field is considered for the sake of simplicity) in view of the fact that the present section does not offer a complete derivation of (7-11). This principle, which follows as a trivial first integral (with respect to time and space variables) of the equation of motion, expresses the principle of conservation of energy of a moving volume element of the fluid, as explained below.

7.3.6 Energy conservation: Bernoulli's principle

In numerous situations of practical interest, one may ignore *dissipative* effects in the fluid like, for instance, viscosity, where energy is lost into the body of the fluid due to an internal friction. If, in addition, the flow happens to be a steady one, then one may apply the principle of conservation of mechanical energy to any small portion of the fluid as it moves along a stream line (refer to the last paragraph of sec. 7.3.5).

In the course of such motion the kinetic energy of the small portion under consideration, of mass, say, δm , gets changed due to the external forces acting on the fluid particles and performing work on the small portion, where the term 'external' refers to forces originating outside the body of the fluid. If the external forces are conservative in nature then the work done by these forces can be expressed as the decrease in potential energy of the portion of fluid under consideration. Let us assume that the only external force acting on the fluid particles is that of gravity, in which case the change in gravitational potential energy of the small portion between any two given time instants is $\delta mg(h_2 - h_1)$, where h_1, h_2 are the heights of the points above some fixed reference level at which this portion of the fluid is located at those two time instants (see fig. 7-17). This change in potential energy is accompanied by an increase in kinetic energy, in accordance with the principle of conservation of energy.

In addition, there occurs a further change in the kinetic energy since work is done on

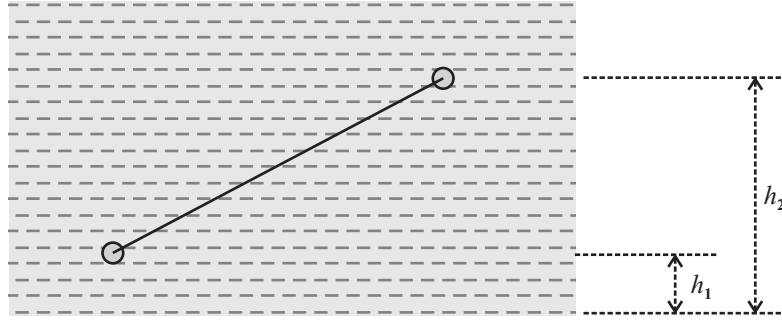


Figure 7-17: Locations of a small volume element of a fluid at two time instants; the element moves along a stream line; heights above a fixed reference level at the initial and final instants are h_1 and h_2 ; the terms 'initial' and 'final' refer to the beginning and end of any chosen time interval.

the small mass δm as it moves across other similar portions of the fluid, by the internal stress forces due to these other mass elements. As already mentioned, these stress forces are entirely described at any given location by the pressure in the fluid. For an *incompressible* liquid, the work done on the mass element δm as it moves from one point to another happens to be $(-\frac{\delta m(p_2 - p_1)}{\rho})$, where p_1 , p_2 are the pressures at the initial and final positions, and ρ is the density of the fluid. Notice that this expression for the work done by the internal forces can also be looked upon as a change in an appropriately defined 'potential energy', which means that the internal forces are of a conservative nature. This is so because we have assumed dissipative effects to be negligible.

The principle of conservation of energy demands that the total increase in kinetic energy of the mass element as it moves from one point to another must be equal to the work done on it by the internal and external forces taken together, the work by the external forces being given by the decrease in its potential energy. The expression for the increase in kinetic energy is $\frac{1}{2}\delta m(v_2^2 - v_1^2)$, where v_1 , v_2 denote the initial and final velocities of the element under consideration. In other words, one has the following equality describing the conservation of energy during the motion of the small mass element of the fluid during the course of its motion along a stream line from one point to another:

$$h_1 + \frac{v_1^2}{2g} + \frac{p_1}{\rho g} = h_2 + \frac{v_2^2}{2g} + \frac{p_2}{\rho g}. \quad (7-12)$$

(check this relation out; the total work done on the mass element by gravitational force

and by the stress force equals the change in kinetic energy).

Put differently, the following quantity remains constant during the motion of the small mass element under consideration:

$$h + \frac{v^2}{2g} + \frac{p}{\rho g} = \text{constant}. \quad (7-13)$$

The equations (7-12) and (7-13) are equivalent expressions for Bernoulli's principle for an incompressible fluid.

While I have written down the equation with reference to the steady flow of an incompressible liquid, a more general form of the equation can be arrived at for the steady flow of an ideal fluid that need not be incompressible, where dissipation effects can be ignored. Moreover, the general form is applicable even when conservative external forces other than the force of gravity act on the fluid.

At any given point on a stream line, the three terms in the expression (7-13) are referred to as the *gravity head*, the *velocity head*, and the *pressure head* respectively (a more general expression for the pressure head is to be employed if the fluid is not incompressible). So one way to state Bernoulli's principle is to say that in the steady flow of a fluid, the sum of the above three *heads* remains constant along any given stream line. For the sake of brevity we denote the three heads as H_G , H_V , and H_P respectively. One then has

$$H_G + H_V + H_P = \text{constant (on a stream line)}. \quad (7-14)$$

If the liquid is in equilibrium then the velocity head $H_V (= \frac{v^2}{2g})$ is zero everywhere and then (7-12) or (7-13) simply expresses the variation of pressure with height in a liquid at rest as expressed in (7-2) (check this out) and tells us that the pressure in a liquid at rest decreases in proportion to the height above any given reference level and increases in proportion to the depth below the level. For a liquid in motion, however, changes in H_P and H_G need not be equal and opposite since H_V now enters into the picture and the

total head, made up of the three taken together, remains constant along a stream line.

As stated here, Bernoulli's principle applies to the steady flow of an ideal fluid (we have, in addition, assumed the fluid to be incompressible for the sake of simplicity), regardless of whether the flow is rotational or irrotational. A great simplification occurs in the case of an irrotational flow since in that case the sum of the three heads in (7-14) turns out to be a constant, not only on a stream line, but *throughout the volume of the flow*, i.e., the value of the sum does not vary from one stream line to another.

7.3.7 Potential flow

The steady irrotational flow of an ideal incompressible fluid is referred to as a *potential flow* since the velocity field for a such a flow is completely determined by a scalar function of position $\phi(\mathbf{r})$, namely, the *velocity potential* introduced in sec. 7.3.3 where, in virtue of formula (7-10b), the potential distribution in space can be determined in a manner analogous to the determination of the potential distribution in an *electrostatic problem* in a region of space devoid of charges. Though defined in terms of a set of ideal conditions, potential flows are often invoked as simple models, starting from which a number of features of flows under more realistic conditions can be understood.

Fig. 7-18 depicts the configuration of stream lines in the case of a potential flow past a circular cylinder in the absence of an external force field, in a plane perpendicular to the axis of the cylinder. Knowing the velocity V of the fluid at an infinitely large distance from the cylinder axis, one can solve for the potential, and hence for the velocity field, at all points, thereby obtaining the stream lines as shown in the figure. One observes that the distribution of stream lines is symmetric between the 'front' and 'rear' sides of the flow (i.e., upstream and downstream) and, as the Bernoulli principle implies, so is the distribution of pressure. The circular symmetry of the cylinder also implies that the pressure is symmetrically distributed above and below the line XOX' through the center (the two sides are marked as 'top' and 'bottom' in the figure). In other words, the cylinder experiences no net force due the flow around it.

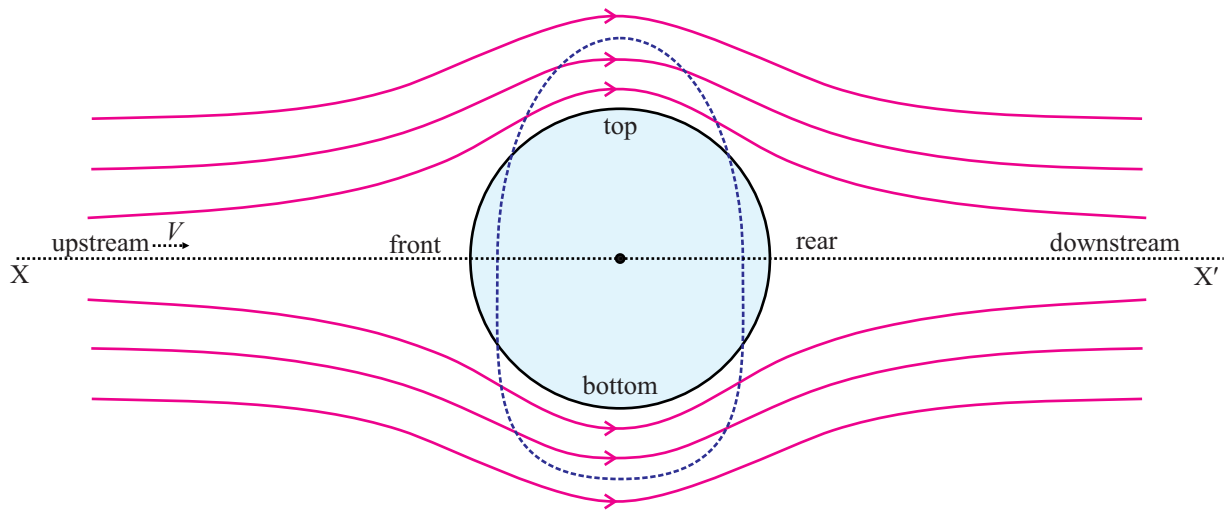


Figure 7-18: Depicting the stream lines, for a potential flow past a long circular cylinder, in a plane perpendicular to its axis; XOX' is a line through the center of the circular cross-section, parallel to the flow velocity V at a large distance from the center O ; the front (upstream), rear (downstream), top, and bottom sides are indicated; the cylinder experiences no net force; for the sake of comparison, an asymmetric body is also shown (by dotted line), for which the top and bottom surfaces are of different shapes, resulting in a *lift* force acting from the side of the flat bottom surface to the curved top surface.

7.3.8 Lift and drag forces: a brief introduction

One can compare the potential flow past a circular cylinder with one past an asymmetric body (with cross-section shown by dotted line in fig. 7-18) for which the shapes of the top and the bottom surfaces are different. In this case, an application of Bernoulli's principle shows that the pressure on the curved top surface is less than that on the flat bottom surface, as a result of which there arises a net *lift* force on the cylinder in the direction from the bottom to the top. A similar lift force arises in the case of the flow past a *rotating* cylinder.

The potential flow past a circular cylinder differs from the flow of a real fluid (one with a non-zero viscosity) in one important respect, namely, the formation of a *boundary layer* (see sec. 7.5.8) in the case of a viscous flow, there being no boundary layer for an ideal flow. The boundary layer is caused by the fact that a viscous fluid (even one with a low viscosity) tends to stick to the surface of a solid body with which it is in contact, and results in a *drag* force acting on the body. The fact that a real fluid sticks to the

surface of a solid body with which it is in contact, is an experimentally observed one, and is referred to as the *no-slip* condition. It is made use of in determining the flow characteristics of a fluid under a given disposition of boundary surfaces by demanding that the flow is to be consistent with the equations of motion for the fluid (refer to sec. 7.5.3).

While we have considered here the lift and drag forces exerted on a body by a fluid flowing past it with a stream velocity (i.e., the velocity at upstream points far removed from the body) V , identical forces can be seen to arise if, instead, the body moves with velocity V through a stationary fluid in the opposite direction. The stream lines in this case are obtained by superposing a constant velocity $-V$ on the velocity field obtained for the fluid flowing past the stationary body.

Lift and drag forces exerted by a fluid flowing past a solid body are more generally described in terms of the distribution of the *stress force* over the surface of the body, where the pressure and viscous forces appear as the normal and tangential components of the stress force. In the above simplified introduction, the lift force has been associated with the pressure distribution that can appear even for a non-viscous fluid, and the drag force with the distribution of tangential stress of viscous origin. More generally, both the normal and tangential forces contribute to the lift and *also* to the drag, where the two are defined as resultant forces *perpendicular* and *parallel* to the velocity of relative motion (between the fluid and the body under consideration) respectively, considered at points away from the body where the relative velocity is assumed to be uniform.

7.4 The siphon

I will now present a simplified explanation of the working principles of a *siphon* as an illustration of Bernoulli's principle at work.

Figure 7-19 shows an inverted U-shaped tube with one end (A) dipped in a liquid in a reservoir (V) and with the other end (B) at a level lower than the liquid level in the reservoir. When the arms of the inverted U-tube are filled with the liquid, with no air

bubble or any other discontinuity in the body of the liquid filling the tube, it is found that the liquid from the reservoir spills out of the end B, *climbing up* the arm dipped in the vessel up to the level marked C. The process continues till the liquid in the reservoir is exhausted or the liquid level in it dips below the end B.

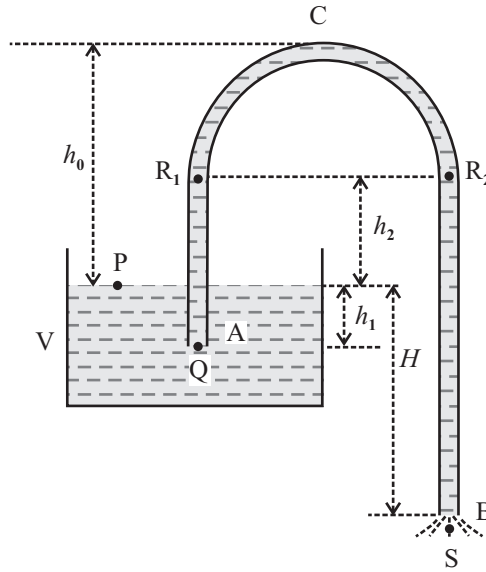


Figure 7-19: Illustrating the working principle of a siphon; liquid is drained from the vessel V by means of the inverted U-shaped tube, one end of which (A) is dipped in it, with the other end (B) at a level lower than the liquid level in V; the liquid spills through B, climbing up the height h_0 of the ascending part of the tube; Q, R_1 , R_2 , and S are points on a stream line (see sec. 7.3.6); the velocity of efflux through B increases with H , till it reaches a certain maximum value determined by h_0 ; in turn, h_0 has to be less than a certain maximum value for the siphon to work under ordinary conditions; h_0 can, however, be increased beyond this under specially maintained conditions.

This set-up for removing liquid from a vessel where the liquid has to clear a height (up to the level C in the figure) before spilling out, is called a *siphon*. In some instances, the end B may be dipped in liquid in another vessel (fig. 7-20), the liquid level in this second vessel being lower than that in the reservoir to be drained. The principle underlying the working of the siphon is found to be at work in a large number of appliances and natural settings.

While the liquid in the arm CB of the siphon tends to fall off due the gravitational pull, liquid from the reservoir rises in the arm AC due to atmospheric pressure and is made to

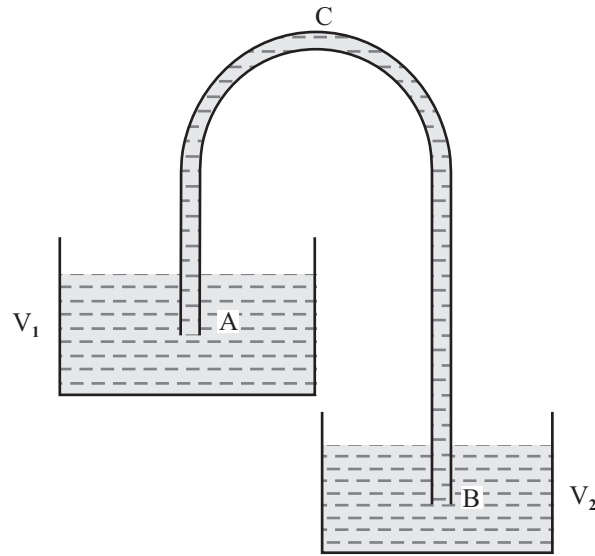


Figure 7-20: Alternative arrangement for siphon; liquid is drained from the vessel V_1 into a second vessel V_2 .

spill into the arm CB (however, other mechanisms are also at work, see below), thereby keeping the siphon filled with the liquid and causing the flow - up through the arm AC and down through CB - to continue.

In outlining the working principle of a siphon, we assume for the sake of simplicity that the liquid under consideration is incompressible and that the flow is a steady and ideal one. A simple explanation then follows from the application of Bernoulli's principle to the flow.

Imagine a stream line passing close to the points P, Q, R_1 , R_2 , S (fig. 7-19), of which P lies on the surface of the liquid, far from the siphon inlet, in the vessel (V) to be drained, and the remaining are points in the flow tube ACB, Q and S being located at the two ends and R_1 , R_2 being at the same horizontal level near the top as shown. Assuming the tube to be a sufficiently narrow one (but not too narrow to cause appreciable break-up of the stream lines and to result in a lowering of the working efficiency of the siphon), one can expect such a stream line to exist. Note that the heights, measured from the horizontal level at S, of the points mentioned above are respectively H , $H - h_1$, $H + h_2$, $H + h_2$, 0, where H is the height, above S, of the water level in the vessel V, h_1 is the depth of

the siphon inlet below this level, and h_2 is the height of R_1 , R_2 above the same level, as shown in the figure.

The pressures at P and S may both be assumed to be the atmospheric pressure (say, P_0), since both are exposed to the atmosphere. On the face of it, then, it might appear that the pressures at the points R_1 , R_2 , located in the same horizontal level just below the top (C) of the siphon tube are respectively $p_{R_1} = P_0 - h_2\rho g$, and $p_{R_2} = P_0 - (H + h_2)\rho g$, i.e., $p_{R_1} - p_{R_2} = H\rho g$, and that it is *this* pressure difference that drives the liquid from R_1 towards R_2 . On the basis of such an explanation one would expect that the liquid gains in velocity as it moves through the siphon tube.

In reality, however, this is not a correct description of what happens inside the siphon during the steady flow of the liquid through it. Indeed, assuming the cross-section of the siphon tube to be uniform, the velocity has to be the *same* throughout the tube for steady flow to be possible, and the liquid *cannot* gain in velocity anywhere within the tube.

Problem 7-5

Show that the velocity of steady flow an incompressible fluid through a tube of uniform cross section must also be uniform.

Answer to Problem 7-5

HINT: Assuming the cross-section at any point in the tube to be a and the velocity of the liquid past this point to be v , show that the volume of liquid flowing through this cross-section per unit time is av ; since the liquid is incompressible and since in steady flow there cannot be any accumulation of liquid anywhere in the tube, show that this implies the above statement. This constitutes a particular instance of the equation of continuity (refer to sec. 7.3.4) for the steady flow of an incompressible fluid. Here we assume that the velocity is the same at all points on a cross section of the tube. This is a valid assumption for a liquid of low *viscosity* flowing through a tube of sufficiently large cross section, where the *boundary layer* formed on the wall of the tube is of a negligible thickness. Refer to sec. 7.5 for an introduction to the relevant concepts.

To be more precise, the error in the above naive argument lies in the assumed values of p_{R_1} , and p_{R_2} , where the *velocity head* has not been taken into account. A more correct approach would be as follows.

Considering the stream line through P, Q, R_1 , R_2 , S mentioned above, one can invoke Bernoulli's principle, expressed by (7-12) for the points P and S. Assuming the vessel V to be a wide one, the liquid surface in it may be assumed to be almost stationary as the liquid is drained through the narrow siphon tube, i.e., the velocity head at P may be assumed to be zero. Further, since the surface is exposed to the atmosphere, the pressure at P will be P_0 , the atmospheric pressure. This may also be assumed to be the pressure at S, the outlet end of the siphon, which is similarly exposed to the atmosphere. Finally, noting that the gravity head at P, with the horizontal plane through S taken as the reference level, is H , and assuming the speed of efflux through the outlet end to be v , we have

$$\frac{P_0}{\rho g} + H = \frac{P_0}{\rho g} + \frac{v^2}{2g}, \quad (7-15)$$

from which we get the speed of efflux v :

$$v = \sqrt{2gH}. \quad (7-16)$$

In other words, the velocity of efflux is the same as that of free fall through a height H , the depth of the siphon outlet below the liquid level in V. This must also be the velocity (we use the term 'velocity' loosely, in the sense of speed) at every point of the siphon tube.

Consider next the same principle applied to R_1 and S or to R_2 and S. As discussed above, the velocity at both R_1 and R_2 must be v . The gravity head at either of these two points is $(H + h_2)$ (fig. 7-19), and hence the pressure at the two points must also be the same, say p . This gives

$$H + h_2 + \frac{v^2}{2g} + \frac{p}{\rho g} = \frac{P_0}{\rho g} + \frac{v^2}{2g}, \quad (7-17)$$

from which we obtain the pressure at each of the points R_1 , R_2 :

$$p = P_0 - (H + h_2)\rho g. \quad (7-18)$$

Note how this result differs from the naive conclusion, indicated above, without properly taking into account the energy conservation principle expressed through (7-12), (7-13).

Finally, applying (7-12) to the siphon inlet Q and outlet S , the pressure (p_Q) at Q works out to

$$p_Q = P_0 - (H - h_1)\rho g, \quad (7-19)$$

which shows that the pressure difference between Q and R_1 (or R_2) is completely accounted for by the difference between the gravity heads as it must be, since the velocity heads at these two points are the same (compare with the pressure difference between P and R_1).

On the face of it, it might appear from (7-16) that the velocity of efflux, and hence the rate of draining (av) by the siphon may be increased indefinitely by increasing H . That this is not ordinarily the case may be seen as follows. Applying (7-12) to the points C (topmost point in the siphon tube) and S one finds the pressure at C to be

$$p_C = P_0 - (H + h_0)\rho g \quad (7-20)$$

where h_0 is the height of C above P , i.e., *the height of ascent* in the siphon tube (fig. 7-19).

(Check this out.)

Since the minimum possible value for p_C is zero (however, see below) one obtains the

maximum possible value of H , and hence of v , from (7-20):

$$v_{\max}^2 = 2\left(\frac{P_0}{\rho} - h_0 g\right). \quad (7-21)$$

This result tells us that the maximum velocity of efflux in a siphon decreases as h_0 , the height of ascent in the siphon tube increases. Note that v_{\max} becomes zero for $P_0 = h_0 \rho g$. This gives the limiting value of the height of ascent ($h_{0\max} = \frac{P_0}{\rho g}$) beyond which a siphon would ordinarily cease to work since otherwise the pressure at the topmost point C would have to be *negative* to support the flow through the siphon. It is in this sense that the siphon can be said to be a device driven, under ordinary circumstances, by the atmospheric pressure since the latter has to be adequate to support a column of liquid of height h_0 .

Two comments are, however, in order. First, our deductions above are valid only under assumption that the system is conservative, i.e., there is no energy *dissipation* in the system. In reality, the velocity of efflux does not reach the predicted value (equation (7-16)) due to various types of energy loss. And secondly, the siphon may be made to work even when $h_0 \rho g > P_0$ by means of a number of *special measures*. These include using a siphon tube with smooth and clean surface, and a liquid with no dissolved gas in it so that *cohesive forces* between the liquid molecules can maintain the flow *even at a negative pressure* (say, at C). Indeed, with such special measures, *a siphon may be made to work even in a chamber from which the air has been drawn out*.

Problem 7-6

The efflux point of a siphon is at a vertical depth of $H = 2.0$ m below the water level in the vessel to be drained with its help, and $l = 3.5$ m below the highest point to which the liquid rises in the siphon tube. What is the velocity of efflux? If the area of cross-section of the siphon tube is $\alpha = 2.5 \times 10^{-5}$ m², what is the volume of water drained per second? What is the maximum possible velocity of efflux that can be attained by lengthening the descending arm of the tube? (Acceleration due to gravity: 9.8 m·s⁻², density of water: 1.0×10^3 kg·m⁻³, atmospheric pressure: 1.01×10^5 Pa.

Answer to Problem 7-6

HINT: The velocity of efflux is $v = \sqrt{2gH} = \sqrt{2 \times 9.8 \times 2.0} \text{ m}\cdot\text{s}^{-1}$. The volume of water drained per second is $V = \alpha v = 2.5 \times 10^{-5} \times \sqrt{2 \times 9.8 \times 2.0} \text{ m}^3\cdot\text{s}^{-1}$. The maximum possible velocity of efflux is $v_{\max} = \sqrt{2(\frac{P_0}{\rho} - (l - H)g)} = \sqrt{2(\frac{1.01 \times 10^5}{1.0 \times 10^3} - 1.5 \times 9.8)} \text{ m}\cdot\text{s}^{-1}$. Here the height of ascent in the rising arm of the siphon is $h_0 = l - H$.

7.5 Viscosity and fluid flow

7.5.1 Introduction

Imagine a fluid in motion, in which the velocities of fluid particles in successive layers of the fluid differ from one another (see fig. 7-21 by way of illustration), i.e., the successive layers are in *relative motion*. It is an observed fact that, in such a situation, contiguous layers in the fluid exert a force on one another that tends to diminish the velocity differential. In other words, faster layers are slowed down while the relatively slower ones are speeded up. In this, these internal forces in a fluid resemble the forces of friction commonly observed between bodies in contact having a relative motion with respect to one another. This phenomenon, resembling friction, of internal forces being brought into play tending to diminish the relative motion between successive layers in a moving fluid, is termed *viscosity*. Every fluid is characterized by a certain *coefficient of viscosity* (at times referred to as its *viscosity* for the sake of brevity), defined as in sec. 7.5.2 below. The consideration of the effects of viscosity removes the restrictive assumption of ideal flow made in earlier sections of this chapter.

7.5.2 Newton's formula for viscous force

The coefficient of viscosity of a fluid is defined with reference to *Newton's formula* that expresses a basic fact describing the viscous force generated in the fluid.

Figure 7-21 depicts schematically a fluid in motion (in the direction of the double-headed arrow) in which a pair of parallel planes (C_1 , C_2) are imagined such that the fluid particles in each of these planes share a common velocity, different from the velocity of the

particles in the other. Let the planes be perpendicular to the z -axis of a Cartesian co-ordinate system, with co-ordinates, respectively, z and $z + \delta z$, and let the corresponding velocities of the fluid particles in the planes (say, along the x -axis of the co-ordinate system) be v and $v + \delta v$, where δv and δz will be assumed to be infinitesimally small quantities.

Looking at the motion of the fluid particles in the plane C_1 , the rate of change of velocity with distance along the z -axis, is then seen to be $\frac{\delta v}{\delta z}$ which, in the limit of δz and δv being infinitesimally small, can be written as $\frac{dv}{dz}$, and is termed the *velocity gradient* at the layer C_1 with co-ordinate z .

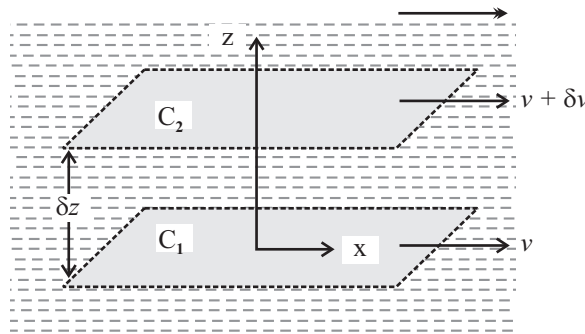


Figure 7-21: Illustrating the production of viscous force in a moving fluid; the x - and z -axes of a Cartesian co-ordinate system are chosen as shown; C_1 and C_2 are two layers with co-ordinates z and $z + \delta z$, the velocity of fluid particles in these two being v and $v + \delta v$ along the x -axis; considering the limit $\delta z \rightarrow 0$, the velocity gradient at C_1 is seen to be $\frac{dv}{dz}$; the liquid above the layer C_1 then exerts a viscous force on the liquid below it, given by the formula (7-22); contiguous layers tend to impede the relative motion between one another.

The basic idea underlying Newton's formula is that the portion of the fluid above the plane C_1 exerts a force on the portion below C_1 because of the velocity gradient existing within the fluid at the layer under consideration. Considering any area, say A , in C_1 , the fluid immediately above this area exerts a force F on the fluid immediately below it, where this force, acting along the x -direction, is proportional to A as also to the velocity gradient $\frac{dv}{dz}$. Moreover, this force acts in the positive x -direction if the velocity gradient is positive, i.e., if the velocity along the x -direction increases upward, because in that case the fluid below C_1 tends to be accelerated by the force exerted by the fluid above it.

In other words, one can write

$$F = \eta A \frac{dv}{dz}, \quad (7-22)$$

where η is a constant characterizing the fluid under consideration. It is this constant that is referred to as the coefficient of viscosity of the fluid.

1. By contrast, the force exerted by the fluid below C_1 on that above it is given by the expression

$$F = -\eta A \frac{dv}{dz}, \quad (7-23)$$

and tends to retard the motion of this part of the fluid if $\frac{dv}{dz}$ is positive.

2. The viscous force has been interpreted above as the force exerted by the liquid immediately above the layer C_1 on the liquid immediately below it (or *vice versa*) through the surface of separation C_1 . This, at times, is interpreted as the force between the two layers C_1 and C_2 in the limit $\delta z \rightarrow 0$.

The unit of the coefficient of viscosity defined as above is $\text{N}\cdot\text{s}\cdot\text{m}^{-2}$ or, equivalently, $\text{Pa}\cdot\text{s}$. The viscosity of a fluid varies to a considerable degree with its temperature. The viscosity of water at 300 K is $8.6 \times 10^{-4} \text{ Pa}\cdot\text{s}$.

Problem 7-7

A block with a smooth surface is made to move uniformly over a film of liquid on another smooth horizontal surface. The surface area of the block in contact with the liquid is 0.01 m^2 while its velocity is $1.0 \text{ m}\cdot\text{s}^{-1}$. If the thickness of the film be 0.002 m and the coefficient of viscosity of the liquid be $8.0 \times 10^{-4} \text{ Pa}\cdot\text{s}$, estimate the force required to maintain the motion of the block.

Answer to Problem 7-7

If the block is to move with a uniform velocity, the force applied to it has to be balanced by the opposing force exerted by the film of liquid which, in turn, is equal in magnitude to the viscous force exerted by a thin layer of liquid, in contact with the block, on the film below it (reason this

out). We can assume that the velocity gradient in the liquid film perpendicular to the direction of motion is $\frac{dv}{dz} = \frac{1.0}{0.002} \text{ s}^{-1}$. One can now invoke formula (7-22) to obtain an estimate for the force as $F = 8.0 \times 10^{-4} \times 0.01 \times \frac{1.0}{0.002} \text{ N}$, i.e., $4.0 \times 10^{-3} \text{ N}$.

7.5.2.1 Kinematic viscosity

Incidentally, in the theoretical analysis of fluid motions it is often found that, instead of the coefficient of viscosity (η), a quantity of more direct relevance is the *kinematic viscosity* (ν) defined as

$$\nu = \frac{\eta}{\rho}, \quad (7-24)$$

where ρ stands for the density of the fluid under consideration. The unit of kinematic viscosity is $\text{m}^2 \cdot \text{s}^{-1}$.

In order to distinguish it from the kinematic viscosity, the coefficient of viscosity of a fluid is sometimes referred to as the *dynamic viscosity*.

7.5.2.2 Variation of viscosity with temperature

The coefficient of viscosity of a fluid depends on the temperature. A fact of basic relevance in this context is that *the nature of temperature variation of the viscosity of liquids differs fundamentally from that of gases*. While the viscosity of a gas increases with an increase in the temperature, that of a liquid *decreases*, and that too more rapidly compared to the rate of temperature variation for a gas. In other words, $\frac{d\eta}{dT}$, which is negative for a liquid and positive for a gas, is usually larger in magnitude for a liquid as compared to a gas.

In this context, highly compressed gases are found to resemble liquids in that their viscosity decreases with a rise in temperature. On the other hand, there are a few liquids

(e.g., liquid helium and liquid sulphur) that show a positive temperature coefficient of viscosity over certain ranges of temperature.

This difference in the nature of temperature variation of the coefficient of viscosity reflects a fundamental distinction between liquids and gases in respect of the origin of the viscous force, which we will have a brief look at in sec. 7.5.7.

7.5.3 Viscosity and transport of momentum

Considering any point P in a fluid in motion, and imagining a planar area around that point, an issue of considerable relevance relates to how the fluid on one side of the area affects the fluid on the other side (see fig. 7-22). Recall that essentially the same question was involved in our earlier considerations on the state of stress at a point in a deformable body (refer to sec. 6.4) and the pressure at a point in a fluid (sec. 7.2.1). While, in these two earlier contexts we talked of the *force* exerted by one part of the deformable body or the fluid on another through a common boundary, we will now change our language a bit and, instead, will talk of the *transport of momentum*.

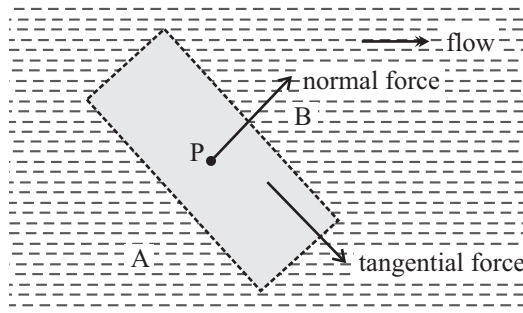


Figure 7-22: Two portions (A and B) of a fluid in motion with a common surface of separation around a given point P in it, where a transport of momentum occurs through this common surface, resulting in the development of stress forces; compare with a similar figure (fig. 7-1) for a fluid at rest; for a fluid in motion, the stress force exerted by A on B has not only a normal component, corresponding to the pressure at P, but a tangential component as well.

Indeed, the force exerted by one part of a fluid (say, A) on another part (say, B) through a common surface of separation between the two can be described as a transport of

momentum from the former to the latter through the common surface, where the force is the rate at which momentum is transported. For any given point P and any given orientation of the planar surface under consideration, the way the two parts of the fluid affect each other is completely described by a set of quantities referred to as the components of the *stress tensor* at the point P. In the case of a fluid at rest, the stress tensor reduces to a scalar, namely the pressure at the point under consideration (with a negative sign) since, as we saw in sec. 7.2.1, the force exerted by A on B per unit area of the common surface of separation between them happens to be independent of the orientation of the latter, and is always perpendicular to it, being directed from A toward B.

For a fluid in motion, on the other hand, the momentum transfer between the two parts under consideration is described not by one but, in general, by *three* independent quantities, namely, the pressure (p), the coefficient of viscosity (η), and a *second* viscosity coefficient (ζ). Of the two viscosity coefficients η and ζ , the latter is referred to as *bulk* viscosity while the former is sometimes called the *shear* viscosity in order to distinguish it from the bulk viscosity.

However, the bulk viscosity is not relevant in describing the motion of an incompressible fluid. Since the compressibility of a liquid is ordinarily very low, the motion of liquids is adequately described in terms of only one single viscosity coefficient, namely, the shear (or *dynamic*) viscosity (η) or, equivalently, by the kinematic viscosity (ν). The bulk viscosity assumes relevance in situations involving rapid expansion or contraction of volume elements within a liquid.

Notice from fig. 7-21 and eq. (7-22) that the shear viscosity can be looked upon as a resistance to shearing motion, i.e., a motion in which successive fluid layers tend to slide over one another. Thus, when an external tangential force is applied on a fluid layer tending to make it slide over a contiguous layer, the fluid starts flowing since *a fluid cannot withstand shear while in equilibrium*, and at the same time, an internal force resembling a frictional resistance is brought into play, described by eq. (7-22).

Imagining a small volume element within a fluid, one can write out its equation of motion by equating its rate of change of momentum to the total force acting on it, where the total force includes the external as well as the internal forces on the element, the latter being made up of all the stress forces arising due to momentum transfer from contiguous parts of the fluid to the element under consideration. For an incompressible fluid, these stress forces involve the pressure acting normally through the boundary surface of the element as well as the viscous force described by Newton's formula. The resulting equation of motion is referred to as the *Navier-Stokes* equation for fluid flow. For a compressible fluid, the Navier-Stokes equation involves the bulk viscosity (ζ) in addition to the shear viscosity (η), and constitutes the generalization of the equation of motion (formula (7-11)) of an ideal fluid.

Newton's formula (7-22) expresses a simple linear relation between the rate of production of shearing strain and the shearing stress generated within a liquid, which can be generalized to a linear relation between the stress tensor and the rate of change of the strain in which the pressure, the shear viscosity, and the bulk viscosity make their appearance. Such a linear relation, however, is not of general validity and there occur a large class of situations in which the relation between the stress tensor and the rate of change of strain is more complex. This will be discussed in brief outline in sec. 7.5.5.

7.5.4 Viscosity and turbulence

The type of fluid motion depicted in fig. 7-21 is an instance of *laminar flow* briefly introduced in section 7.3.1. In a laminar flow, the instantaneous motion of fluid particles is directed along stream lines, which are well-defined curves of regular shape. Though these curves may change in shape and in their position in space in the course of time, the motion locally resembles a sliding of fluid layers past one another.

By contrast, a *turbulent* flow (refer to sections 7.3.2, 7.5.9) involves an irregular and unpredictable motion of the fluid particles, where stream lines lose their integrity, and the picture of layers sliding past one another does not apply for any given length of time.

While the coefficient of viscosity of a fluid has been defined by referring to a regular, laminar flow, it retains significance in the description of turbulent flows as well. Indeed, a turbulent flow involves a complicated mixture of regular and irregular features. A number of features in a turbulent flow are, in the main, independent of the viscosity of the fluid under consideration while a number of other features, notably the ones at small length scales, depend on the viscosity. For instance, the *dissipation* of energy in a turbulent flow, i.e., the conversion of kinetic energy of motion into internal energy that ultimately spreads out through the bulk of the fluid, depends on the viscosity, as in a laminar flow (refer to sec. 7.5.8). In other words, the viscosity of the fluid retains a limited, though crucial, significance in a turbulent flow. The Navier-Stokes equation continues to be the equation of motion of the fluid, though its relevance in arriving at a detailed prediction of the fluid flow from a given set of initial conditions gets drastically reduced. To the extent that the coefficient of viscosity retains its relevance in the description of features of a turbulent flow, Newton's formula also remains relevant, where one only needs to specify the derivatives of the velocity field at any point in space and any given instant of time.

7.5.5 Non-Newtonian fluids

However, this entire description of fluid motion in terms of laminar and turbulent flows, and the reference to Newton's formula and the Navier-Stokes equation in describing the flow, applies only to a certain class of fluids referred to as the *Newtonian* ones. There remain, on the other hand, a large class of *non-Newtonian* fluids for which a broader theoretical basis than the one provided by Newton's formula is necessary in the description of their motion.

If one tries to relate the viscous shear stress ($\tau = \frac{F}{A}$) to the velocity gradient ($\frac{dv}{dz}$, also termed the shear rate) in a non-Newtonian fluid by means of an equation of the form (7-22), one will find that the coefficient η is no longer a constant, but depends on the shear rate itself. In other words, the relation between the stress and the shear rate is a *non-linear* one, and η is referred to as the *apparent* viscosity. The variation of the apparent viscosity with the shear rate is shown schematically for three different categories

of non-Newtonian fluids in fig. 7-23, where the linear graph passing through the origin (dotted line), corresponding to a Newtonian fluid, is shown for the sake of comparison.

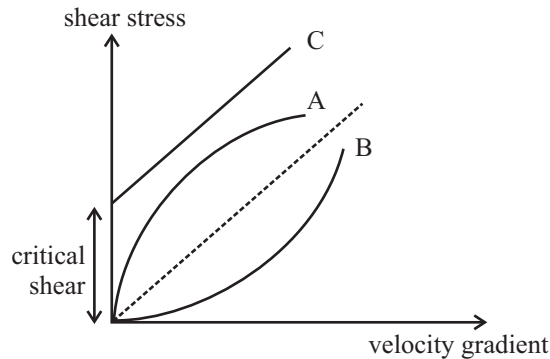


Figure 7-23: Variation of shear stress with velocity gradient for non-Newtonian fluids (A) pseudoplastic, (B) dilatant, (C) Bingham fluid; the case of a Newtonian fluid is shown with a dotted line; the slope of the graph in (A) decreases with the shear rate while that in (B) shows an increase; the graph in (C), though approximately linear, does not pass through the origin.

For the fluids corresponding to the curve marked A in the figure, the apparent viscosity decreases with an increase in the rate of shear (viscous thinning), while for those corresponding to the curve B, the apparent viscosity increases (thickening). The former group, referred to as *pseudoplastic* fluids, includes such materials as paper pulp in water, latex paint and molasses, while an example of the latter, termed dilatants, is a suspension of sand in water (quicksand). A third category (graph marked C in the figure), referred to as *Bingham plastic* (examples: mud, slurry, toothpaste) behaves essentially as a solid for low stresses and flows like a viscous fluid at stresses above a certain critical value.

The class of non-Newtonian fluids also includes those for which the apparent viscosity is *time-dependent*, i.e., there occurs either a 'thinning' or a 'thickening' with an increase in the duration of the shearing flow.

More generally, *complex fluids* are materials that commonly involve a coexistence of more than one phases and have unusual flow properties depending on structures or states of aggregation at an intermediate scale in between the atomic constituents and

the bulk fluid.

The subject of *rheology*, addressing the flow properties of diverse materials, is a vast area of basic and applied science with immense potentialities for practical applications.

7.5.6 Poiseuille's flow

Referring to a Newtonian fluid once again, imagine a fluid flowing through a narrow horizontal cylindrical tube of uniform cross-section, the flow being maintained by a pressure difference between the two ends of the tube. For a sufficiently small pressure difference between the ends, the flow turns out to be a laminar one. Moreover, given sufficient time, the fluid settles down to a *steady* flow when the fluid velocity at every point in the tube remains constant in time. In this steady state, the external forces on any given element of the fluid are balanced by internal stress forces.

Imagining a set of cylindrical surfaces coaxial with the flow tube, each of the surfaces corresponds to a constant fluid velocity as a consequence of the axial symmetry of the problem. Let the fluid velocity on a cylindrical surface of radius r be $v(r)$. The set-up is depicted in fig. 7-24, in which, p_1 and p_2 are the pressures at the two ends of the tube with the pressure difference $p = p_1 - p_2$ causing the fluid to move through it in the direction shown. We assume that the fluid under consideration is an incompressible liquid. Such a steady flow of an incompressible liquid through a tube or a pipe is referred to as *Poiseuille's flow*.

Considering the volume of liquid contained within a cylindrical surface of radius r , the external force on it due to the pressure difference between the two ends of the tube is given by

$$F_{\text{ext}} = \pi p r^2, \quad (7-25a)$$

while the viscous force, exerted by the portion of the liquid outside the cylindrical shell

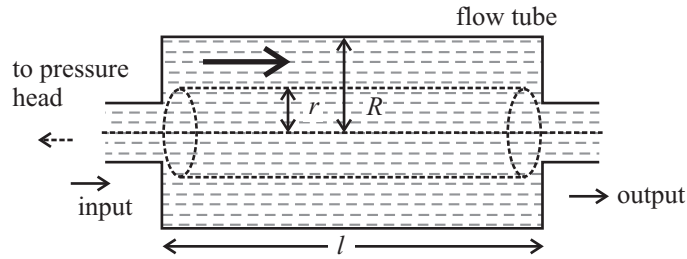


Figure 7-24: Fluid flow through a narrow horizontal tube; the pressure difference driving the flow may be due to a constant head of fluid (say, water) maintained at the input end of the flow tube, whose cross-section is shown in magnification for the sake of clarity; in reality, the length of the tube is taken to be large compared to its radius; the dotted lines depict a cylindrical surface of radius r , for which the fluid contained in its interior at any given instant of time is acted upon by the forces F_{ext} and F_{viscous} , given by equations (7-25a) and (7-25b); the left end face of the cylindrical layer of radius r is taken slightly to the right of the input end.

on the portion in its interior is

$$F_{\text{viscous}} = 2\pi r l \eta \frac{dv}{dr}. \quad (7-25b)$$

Equation (7-25b) is obtained from Newton's formula, eq. (7-22), by putting $A = 2\pi r l$, the area of the cylindrical surface under consideration, where l stands for the length of the flow tube, and noting that the velocity gradient in a direction perpendicular to the velocity reduces in the present context to the rate of change of velocity along the radial direction, i.e., to $\frac{dv}{dr}$.

While the force due to the pressure differential between the two ends has been represented here as an 'external' one on the mass of liquid under consideration, it may, in reality, be a stress force of an internal nature. For instance when one considers the entire mass of liquid at any instant including the liquid in the inlet and outlet ends of the flow tube, the pressures at these two ends are nothing but the normal stresses developed by virtue of the action of one portion of the liquid on another. To be precise, the entrance end of the cylindrical volume containing the mass of liquid on which the forces F_{ext} and F_{viscous} act, is to be taken to be one slightly to the right of the input end (fig. 7-24) of the flow tube since, close to the entrance end, the flow is *non-uniform*, i.e., the flow velocity changes with distance along the axis of the tube, associated with a non-uniform *boundary layer* (see sec. 7.5.8.1).

At the steady state, the sum of the above two forces has to be zero since otherwise the mass of the liquid under consideration would undergo an acceleration. One thus arrives at the relation

$$\frac{dv}{dr} = -\frac{p}{2\eta l}r, \quad (7-26)$$

from which the velocity $v(r)$ at any distance r from the axis is obtained by integration as

$$v = -\frac{p}{4\eta l}r^2 + C, \quad (7-27a)$$

where C is a constant of integration that may be obtained in terms of the value of v for any chosen value of r . Choosing $r = R$, the radius of the flow tube, one can take $v(R) = 0$, since the liquid in contact with the wall of the flow tube may be assumed to be at rest. In other words, one can assume that there occurs no *slip* between the liquid and the solid surface with which it is in contact (refer back to sec. 7.3.8). Making use of this *no-slip condition*, one gets

$$v(r) = \frac{p}{4\eta l}(R^2 - r^2). \quad (7-27b)$$

This is the basic result relating to the steady flow of an incompressible liquid through a cylindrical tube. As a corollary to this equation, the maximum velocity of the liquid in the flow tube, which corresponds to the fluid particles flowing along the axis of the tube, works out to

$$v_{\max} = \frac{p}{4\eta l}R^2. \quad (7-28)$$

One can make use of the relation (7-27b) to work out the rate of flow of the liquid through the tube under the given pressure difference between its ends. The result works out to

$$Q = \frac{\pi p R^4}{8\eta l}, \quad (7-29)$$

where Q stands for the volume of liquid flowing through the tube per unit time (check

this out).

The above expression for Q , referred to as *Poiseuille's formula*, is based on the assumption that the flow in the tube is a steady laminar one. This, in turn, requires that, for a flow tube with a given radius and with a given pressure difference between its ends, the viscosity of the liquid be sufficiently large so that the flow velocity (characterized by v_{\max} or, alternatively, the average flow velocity in the tube) be sufficiently small. Another, more useful, way to express the same condition is that the *Reynolds number* (see sections 1.6.3, 7.5.9) characterizing the flow be sufficiently small.

The Reynolds number for the flow of a fluid of density ρ and viscosity η (kinematic viscosity $\nu = \frac{\eta}{\rho}$) through a tube or pipe of circular cross-section with diameter D is defined as

$$\mathcal{R} = \frac{\rho V_{\text{av}} D}{\eta} = \frac{V_{\text{av}} D}{\nu}, \quad (7-30)$$

where V_{av} stands for the average velocity of flow.

For relatively low values of the Reynolds number (less than 2100 (approx.)), the flow is found to be laminar, while a value larger than 4000 (approx) implies a fully turbulent flow. intermediate values of \mathcal{R} correspond to flows of an intermediate or transitional type where there occurs an apparently random switching between laminar and turbulent flow characteristics. Flows of the transitional and fully turbulent types are common for liquids flowing through pipes of large diameters. Though important from a practical point of view in the context of water and oil distribution systems, the characteristics of such flows are deduced from empirical rather than theoretical considerations.

Problem 7-8

Obtain the Reynolds number for the laminar flow of water (density $\rho = 1000 \text{ kg} \cdot \text{m}^{-3}$, coefficient of viscosity $\eta = 8.0 \times 10^{-4} \text{ Pa} \cdot \text{s}$) through a tube of diameter $1.0 \times 10^{-3} \text{ m}$ and length 0.5 m . If the pressure difference between the two ends of the flow tube be $P = 4.0 \times 10^3 \text{ Pa}$, find the rate of energy dissipation due to the flow. Work out the rate of work done to increase the kinetic energy

of water from its point of injection in the tube to the point where the radial velocity distribution assumes the form (7-27b).

Answer to Problem 7-8

HINT: The Reynolds number is to be calculated from formula (7-30), where the average velocity, worked out from (7-27b) as $V_{av} = \frac{PR^2}{8\eta l}$. Substituting given values, one obtains $\mathcal{R} = \frac{\rho PR^3}{4\eta^2 L} = 390.6$. Assuming the flow to be uniform throughout the length of the tube, the rate of energy dissipation due to viscous resistance is the same as the rate of work done on the water due to the pressure difference, i.e., $W = PQ$ (reason this out), where Q , the rate of flow, is given by formula (7-29). Substituting given values, one obtains $W = 98.1 \times 10^{-5} \text{ J}\cdot\text{s}^{-1}$. Water is injected into the tube at the entrance end with a velocity V_{av} throughout the cross-section, i.e., the kinetic energy injected per second is $K_1 = \frac{1}{2}\pi R^2 \rho V_{av}^3$ (reason this out), where the expression for V_{av} is given above, thus leading to $K_1 = \frac{\pi \rho P^3 R^8}{1024 \eta^3 l^3}$. On the other hand, the kinetic energy of water flowing per second through a thin annular ring of inner and outer radii r and $r + \delta r$ is $\frac{1}{2} \times 2\pi r \delta r \rho v(r)^3$, where $v(r)$ is given by (7-27b) (reason this out). Integrating over the entire cross-section, one obtains the rate of steady flow of kinetic energy through the tube as $K_2 = \frac{\pi \rho P^3 R^8}{512 \eta^3 l^3}$. The difference $K_2 - K_1$ gives the rate of increase of kinetic energy of water within the tube, $K = K_2 - K_1 = \frac{\pi \rho P^3 R^8}{1024 \eta^3 l^3}$. Substituting given values, one obtains $K = 1.2 \times 10^{-5} \text{ J}\cdot\text{s}^{-1}$. This is slightly over 1 per cent of the viscous dissipation.

NOTE: As mentioned earlier, the formula (7-29) is derived by ignoring the non-uniform flow at the entrance end of the flow tube. In the experimental determination of the coefficient of viscosity by the Poiseuille flow method, the formula for viscosity, viz., $\eta = \frac{\pi PR^4}{8QL}$, obtained from (7-29) needs a correction due to the initial non-uniformity of flow, where a first approximation to the corrected formula looks like $\eta = \frac{\pi PR^4}{8QL} (1 - \frac{\rho Q^2}{2\pi^2 PR^4})$. Note that the correction term $\frac{\rho Q^2}{2\pi^2 PR^4}$ is nothing but the ratio $\frac{K}{W}$.

7.5.7 The origin of the viscous force: a brief outline

The viscous force of internal friction in a fluid is, in the ultimate analysis, a stress force originating in the action of one part of the flowing fluid on a contiguous part, where the stress force is a consequence of *momentum transport* from the former to the latter through the common boundary of the two parts. Indeed, the description in terms of a force exerted by one part on another and that in terms of a momentum transfer are

complementary, where one may be more appropriate compared to the other depending on the context.

In the case of a dilute gas, for instance, the intermolecular correlations are almost non-existent, and the forces between the molecules are also negligible, the average distance between them being large compared to the range of the intermolecular forces. In this case, a description in terms of momentum transfer is appropriate since the gas molecules collide continually with one another, and momentum is transferred between these. Considering two parts of the fluid with the fluid particles in one part having a higher velocity compared to the other, there occurs, on the average, a net momentum transfer, from the former to the latter.

In the case of a liquid, on the other hand, a major contribution to the stress force comes from the intermolecular forces between molecules lying close to one another since the intermolecular correlations are larger, and the average intermolecular separation much smaller, in this case. Indeed, in a liquid, the molecules are held in groups or clusters that maintain their identity for short time intervals while on a longer time scale the clusters break up and regroup themselves. Thus, the momentum transfer through intermolecular collisions contributes to the viscous force to a lesser extent compared to a dilute gas. Instead, the momentum transfer is caused by the cumulative effect of the mutual pulls exerted by the clusters on one another.

This distinction between momentum transfer through collisions and that through interactions between molecules in close proximity to one another is, of course, not a fundamental one, since molecular collisions are nothing but one type of interaction between molecules. At the same time, it is not completely devoid of meaning either. At a theoretical level, the stress forces can be worked out, at least in principle, in terms of the energy function (termed the *Hamiltonian*) of the fluid under consideration, which can be expressed as a sum of the kinetic energy function and the mutual potential energy function of the molecules making up the fluid. The distinction between a gas and a liquid depends on the relative importance of these two in determining the momentum transfer rate between contiguous portions of the fluid.

At an increased temperature, the viscosity of a gas increases since the collisions become more frequent, causing momentum to be transported at an enhanced rate. In the case of a liquid, however, the clusters break up more frequently at the higher temperature and the mobility of the molecules increases, causing a reduction in the average pull exerted by the clusters on one another and, consequently, a reduction of the viscosity, where the effect of the increased collision rate is, in general masked. This explains - qualitatively, at least - the difference in the temperature coefficients of viscosity for dilute gases and liquids.

A detailed quantitative analysis for the working out of the stress force is not a simple matter, especially for dense gases and liquids, where definitive results are difficult to arrive at.

7.5.8 The boundary layer

It is of some interest to compare a number of features of the flow of an ideal fluid with those relating to a viscous one, where it is assumed that both of the flows occur under identical conditions (specified in terms of dispositions of boundaries, and of flow parameters at infinitely large distances, where applicable) and that, in the case of the viscous fluid, the coefficient of viscosity is of a sufficiently small value so as to warrant the comparison.

Contrary to intuitive expectations, the two flows turn out to be *qualitatively* different in a number of ways, notable among these being the formation of a *boundary layer* (refer back to sec. 7.3.8) in the case of the viscous fluid since it is an experimentally observed fact that a viscous fluid sticks to the surface of any solid body with which it is in contact, regardless of the value of the coefficient of viscosity. Away from such boundaries, on the other hand, the flow of the real fluid with a low viscosity is qualitatively similar to that of the ideal fluid provided that the flow within the boundary layer is laminar in nature. In the case of *turbulent* boundary layers on the other hand, the effects of turbulence may extend to considerable distances from the boundaries.

Within the boundary layer, there is produced a considerably large velocity gradient, due

to which a *force*, referred to as the *drag force*, is exerted on the solid body on which the boundary layer is formed. In the case of an ideal fluid, the boundary layer vanishes, along with the drag force - the fluid loses its 'bite' on the surface.

In principle, the force exerted on a solid body due to the formation of the boundary layer may include both lift and drag components. Here and in the following, we consider only the latter, for the sake of simplicity.

The motion of the fluid within the boundary layer may be laminar or turbulent in nature or, more generally, one involving both laminar and turbulent regions, with a more or less sharp transition from the former to the latter. The next section (sec. 7.5.8.1) deals with low Reynolds number Poiseuille's flow where the boundary layer is a laminar one, while the transition from laminar to turbulent flow within a boundary layer is illustrated with reference to the flow past a flat plate in sec. 7.5.8.2. The nature of the boundary layer turns out to have a marked effect on the drag force on a body in contact with a fluid in motion (refer to sections 7.3.8 and 7.5.8.3).

7.5.8.1 Laminar boundary layer in Poiseuille's flow

Referring back to sec. 7.5.6 on Poiseuille's flow (more generally, one may consider the flow through a pipe of circular cross-section), an essential simplification in our derivation relates to the assumption that the liquid throughout the entire length of the flow tube is in a state of *uniform* steady motion, with the radial distribution of velocity given by formula (7-27b). In reality, however, the motion near the entrance end of the tube is *non-uniform* where the velocity distribution changes with distance (x) along the axis. For a sufficiently narrow flow tube and a sufficiently viscous liquid, this non-uniformity is taken into account by introducing a small correction wherein the length l appearing in formulae (7-27b) and (7-29) differs from the length of the flow tube, as mentioned in sec. 7.5.6 (in this context, refer to problem 7-8). As in sec. 7.5.6, we assume the flow to be a laminar one.

This non-uniformity of the flow features with distance along the axis relates to the

varying thickness of the *boundary layer* formed near the surface of the flow tube, as shown schematically in fig. 7-25, consistent with the no-slip condition at the surface. Close to the entrance end, the thickness (δ) of the boundary layer is small and it does not fill the entire cross-section of the tube. For increasing values of the distance x , the thickness δ also increases till, at some distance x_0 , δ attains the value R , the radius of the tube. For $x < x_0$, the flow through the central part of the cross-section, not covered by the boundary layer, does not conform to formula (7-27b) and the velocity varies little with the radial distance, i.e., viscosity effects are almost absent. For $x > x_0$, the entire cross-section is covered by the boundary layer and the formula (7-27b) is conformed to (this is commonly referred to as the 'fully developed flow'), where l is to be interpreted as the length of that part of the tube for which the boundary layer fills up the entire cross-section and p as the corresponding pressure drop. Within the fully developed flow, the pressure remains constant in the radial direction at any specified cross-section of the tube, while it decreases along the length of the tube.

The formulae (7-27b) and (7-29) are derived on the assumption that the flow is of the laminar type, corresponding to a relatively low value of the Reynolds number (formula (7-30)). As the Reynolds number is made to increase, there occurs an intermittent regime before the flow becomes fully turbulent.

7.5.8.2 Boundary layer on a flat plate

Fig. 7-26 depicts a cross-section of flow past a flat plate (thick line). A Cartesian coordinate system is shown, with the plate placed in the x-y plane. The fluid velocity relative to the plate is uniform (V) far away from the plate; a cross-section of the boundary layer in the x-z plane is shown, where the thickness (δ) of the layer increases with increasing distance (x) along the length of the plate. The flow within the layer changes from laminar ($x < x_0$) to turbulent ($x > x_0$) at some point x_0 , there being a transition region (not shown) near $x = x_0$ (for such a transition to occur, the plate has to be sufficiently long). The boundary layer is developed symmetrically on the top ($z > 0$) and bottom ($z < 0$) surfaces of the plate.

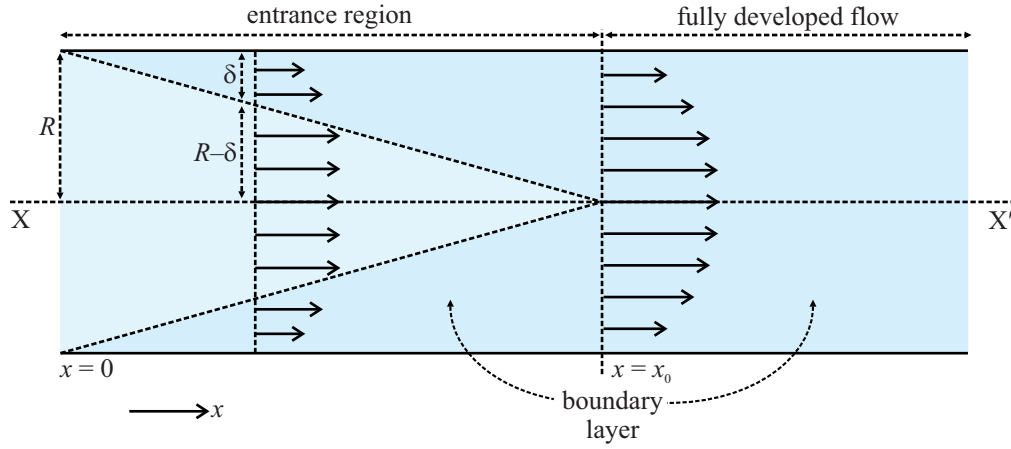


Figure 7-25: Boundary layer (schematic) in Poiseuille's flow (or flow through a pipe of circular cross-section); the thickness (δ) of the boundary layer, formed at the inner surface of the flow tube, is small near the entrance end and increases with the distance (x) along the axis XX' till, at some distance x_0 the boundary layer covers the entire cross section ($\delta = R$, the radius of the tube); for $x > x_0$ the radial distribution of flow velocity (referred to as the 'velocity profile', depicted with arrows of varying length spanning the cross-section) is given by formula (7-27b) while, for $x < x_0$ the velocity profile looks different, with the velocity varying little with radial distance outside the boundary layer ($R - \delta < r < R$, where r stands for the radial distance); inside the boundary layer ($R - \delta < r < R$), the velocity varies more rapidly in order that the no-slip condition at the inner surface of the tube be satisfied; the distance x_0 up to which the boundary layer does not completely cover the cross-section of the tube turns out to be small compared to the length of the tube for a flow with a low value of the Reynolds number.

The Increase in the thickness ($\delta(x)$) of the boundary layer with the distance x from the leading edge, is pronounced near $x = x_0$, the transition point. For any specified value of x , the flow velocity v within the transition layer increases with $|z|$ from 0 at $z = 0$ (the no-slip condition) to approximately the upstream velocity V at $|z| = \frac{\delta(x)}{2}$, the edge of the transition layer at x . The increase of the velocity with $|z|$ is quite sharp at points close to the leading edge while, within the turbulent region ($x > x_0$), the *average* velocity changes more slowly with z (recall that the velocity changes randomly in space and time for a turbulent flow).

In this case of flow past a flat plate, the effective Reynolds number varies from point to point on the plate, its value at any point x being given by the expression

$$\mathcal{R}(x) = \frac{\rho V x}{\eta} = \frac{V x}{\nu}, \quad (7-31)$$

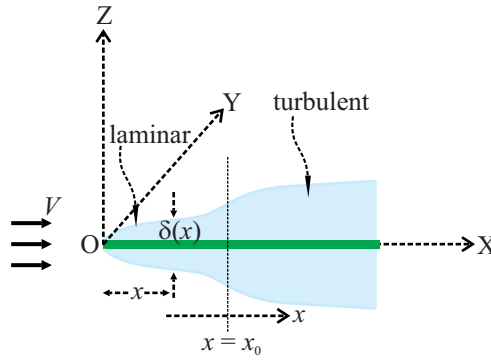


Figure 7-26: Boundary layer (schematic) in flow past a flat plate of negligible thickness; a coordinate system is shown, with the x-axis parallel to the bulk flow and the z-axis perpendicular to the plane of the plate; near the leading edge O, the boundary layer is thin, where the thickness δ increases with distance x along the length of the plate; the Reynolds number \mathcal{R} increases with x , and at $x = x_0$ the flow within the boundary layer changes from a laminar to a turbulent one, there being a transition zone (not shown) near x_0 ; away from the rear edge, the turbulence mixes with the bulk flow; the viscous drag coefficient increases with the degree of roughness of the plate; for a rough plate, the drag coefficient is independent of the upstream bulk velocity V ; the pressure variation is negligible throughout the flow, and there is no pressure drag on the plate; the boundary layer is symmetrical on the two sides of the plate.

and the transition at $x = x_0$ occurs as $\mathcal{R}(x)$ crosses a threshold value which is commonly of the order of 10^5 to 10^6 for smooth plates.

Within the laminar region, the flow characteristics are obtained to a good degree of approximation from a set of differential equations referred to as the *Prandtl equations* that can be derived from the Navier-Stokes equations of motion. In particular, the boundary layer thickness varies as

$$\delta(x) \sim \sqrt{\frac{\nu x}{V}}, \quad (7-32)$$

while the tangential force per unit area on the plate due to the viscous drag is

$$\tau = V^{\frac{3}{2}} \sqrt{\frac{\rho \eta}{x}}. \quad (7-33)$$

In other words, the thickness of the laminar boundary layer increases as \sqrt{x} while the tangential stress decreases inversely with \sqrt{x} . The tangential stress shows an increase near the transition region at $x = x_0$ while, within the turbulent region ($x > x_0$), the drag decreases with increasing distance x . The turbulent drag, however, depends on

the *degree of roughness* of the plate. One can characterize the turbulent flow in terms of a dimensionless *friction coefficient* (or a 'viscous drag coefficient') defined as

$$C_v = \frac{F_v}{\frac{1}{2}\rho V^2 A}, \quad (7-34)$$

where F_v stands for the viscous drag force and A for the area of the plate. This quantity depends, in general, on the Reynolds number as also on the degree of roughness (ϵ) (for a laminar flow, on the other hand, C_v depends on the Reynolds number alone). For a sufficiently high value of ϵ ($> 10^{-2}$, approx) the friction coefficient becomes *independent* of the Reynolds number, corresponding to what is referred to as *fully developed turbulence*. A fully developed turbulence is dominated by *vortices* at various different length scales, with no admixture of regular flow. For lower values of the degree of roughness, there remains a thin *laminar sub-layer* close to the surface of the plate, on top of which the turbulent layer is formed.

In this case of the flow past a flat plate placed parallel to the flow, there is no *pressure component* of the drag force, that arises for the flow of a fluid past a curved obstacle (refer to sec. 7.5.8.3). In the case of a flat plate placed perpendicular to the flow, the pressure drag component is much larger, with the viscous drag being reduced to almost zero value.

As mentioned earlier, the fluid motion within the turbulent boundary layer is of a complex and random nature. The turbulence is made up of 'eddies' of various size, an eddy being a swirling mass of fluid with some size and strength of vorticity of its own (an eddy with zero vorticity, i.e., one involving irrotational flow is also possible as a special case; the fluid rotates as a whole in such a way that there does not occur rotational deformation of any volume element chosen within it).

7.5.8.3 Boundary layer near a curved obstacle: boundary layer separation

In the case of flow past a flat plate (where the upstream flow is assumed to be a uniform one, as in sec. 7.5.8.2), the pressure does not vary appreciably throughout the flow.

This, however, is not the case for the flow past a *curved* obstacle such as a circular cylinder, where a number of new features appear, compared to the flow past a flat plate. In the case of an incompressible liquid with a low viscosity for which the Reynolds number characterizing the flow is high, the boundary layer formed on the surface of the obstacle will be a thin one, outside which the flow differs little from that of an ideal fluid. Assuming for the sake of concreteness that the obstacle is a circular cylinder placed with its axis perpendicular to the flow, which is a uniform one away from the obstacle, one can describe the flow as approximating a potential flow around a circular cylinder, which was briefly introduced in sec. 7.3.7, with the added feature that close to the surface of the cylinder, the flow will be modified due to viscous effects within the boundary layer.

In the absence of the boundary layer (zero viscosity) the stream lines for the potential flow are as in fig. 7-18 for which the pressure at the front surface of the cylinder is distributed symmetrically to that at the rear surface, while the distribution is also symmetric between the upper ('top') and lower ('bottom') surfaces. The velocity distribution is also similarly symmetric, being related to the pressure distribution by Bernoulli's principle.

Considering now the flow of an incompressible fluid with low viscosity, one has to take into account the boundary layer as shown in fig. 7-27, where the variation of velocity along the outer edge of the boundary layer may be taken to be the same as in the non-viscous potential flow, but the velocity distribution within the layer differs markedly due to the non-zero viscosity, the velocity *on the surface of the cylinder being zero* due to the no-slip condition. At the point A at the outer edge of the boundary layer the velocity is zero (the front *stagnation point for the potential flow*), and hence the velocity gradient between A and A' is also zero. In contrast, there appears a considerable velocity gradient between points B and B' ('top'; and similarly between C and C', 'bottom'). The variation of pressure across the boundary layer at any of these points, on the other hand is negligible, implying that the pressure distribution on the front surface of the cylinder is the same as that for the potential flow. The pressure distribution on the *rear* surface, however, may differ because of a lack of front-rear symmetry that characterizes the

viscous flow. This leads to a *pressure drag* on the cylinder that may be small or large depending on the parameters characterizing the flow, to which we will come later.

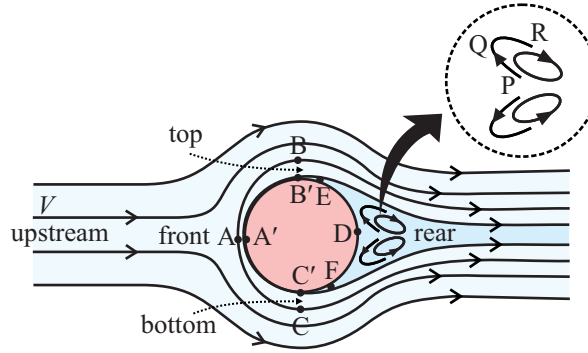


Figure 7-27: Boundary layer (schematic) in flow of a fluid of low viscosity past a circular cylinder with its axis perpendicular to the direction of upstream bulk flow; the front, rear, 'top', and 'bottom' surfaces are marked; A, B and C are points at the outer edge of the boundary layer, while A', B', and C' are corresponding points on the surface of the cylinder, where the velocity is zero; A is the front stagnation point; the pressure difference between A, A' (and also between B, B', C, C') is negligible; however, there is a large velocity gradient across the layer; for instance the velocity changes from zero at B' to approximately the upstream velocity V at B; the flow outside the boundary layer is qualitatively similar to a potential flow, and pressure increases from A, A' to B, B', impeding the acceleration of a fluid element as it moves from A to B; thereafter the velocity of the element decreases to zero at E, even before it reaches the rear end; boundary layer separation occurs at E (and similarly at F on the bottom surface), whereafter there appears a wake, with a back-flow just beyond E, with stream lines curving backward (PQR), and with eddies within the wake; the asymmetry in pressure variation between the front and rear surfaces causes a pressure drag on the cylinder in the direction of flow, which occurs along with the viscous drag due to the tangential stress force exerted by the boundary layer.

Meanwhile, the variation of the velocity gradient across the boundary layer (in a direction perpendicular to the surface) at various points on the surface of the cylinder results in a *friction drag* as in the case of the flow past a flat plate.

In the case of a *spherical* obstacle of radius R the value of the force due to friction drag works out to

$$F_v = 6\pi R\eta V, \quad (7-35)$$

with a notation that is by now familiar. This is referred to as the *Stokes formula* for viscous drag. Strictly speaking, this formula is applicable when the flow does not differ

substantially from the potential flow past a sphere and the boundary layer is not well demarcated from the bulk flow (refer to fig. 7-30(A) below; corrections to the formula become necessary as the Reynolds number is made to increase). Corresponding to the force (7-35), one can define a dimensionless viscous drag coefficient $C_v = \frac{F_v}{\frac{1}{2}\pi\rho R^2 V^2}$, which varies as

$$C_v = \frac{24}{\mathcal{R}}, \quad (7-36a)$$

where the Reynolds number of the flow is given by

$$\mathcal{R} = \frac{\rho V D}{\eta}, \quad (7-36b)$$

$D(= 2R)$ being the diameter of the sphere.

Along with the formation of the boundary layer, there appears the phenomenon of *boundary layer separation* in the case of a viscous flow past a curved obstacle such as the circular cylinder (or the sphere), which is absent in the viscous flow past a flat plate. This results in an asymmetry between the pressure distributions on the front and rear surfaces mentioned earlier, which in turn produces a pressure drag over and above the frictional drag, as we will now see.

Referring to fig. 7-27, the pressure drop from the point A (or A', front) to B (or B', top) causes an acceleration of a small element of the fluid within the boundary layer as it moves from the front to the top. In the case of the potential flow (zero viscosity) this is exactly canceled by the deceleration from the top to the rear (D) where there occurs a symmetrical rise in pressure. For a viscous fluid, on the other hand, there occurs an additional deceleration within the boundary layer all along the surface (from front to top and beyond) due to the non-zero viscosity, and the velocity relative to the surface within the boundary layer gets reduced to zero at some point in between the top and the rear, say, at the point E, which is the point where the separation takes place. The stream lines that lie close to the surface of the obstacle between the points A and E within the boundary layer, get separated from the surface at E, and as one moves beyond E along

the surface, one encounters a *backflow*, with stream lines curving backward (PQR). In between the stream lines that get separated at E and at the corresponding point F towards the bottom, there appears a *wake* involving eddies.

Because of the boundary layer separation, the front-rear symmetry in the pressure distribution of the non-viscous pressure flow is broken, and there develops a pressure deficit at the rear surface of the cylinder compared to the front surface, resulting in the pressure drag mentioned above.

The viscous resistance drag (also referred to as 'skin friction'), the pressure drag (or the 'form drag') and the total drag force depend on a complex of factors, including the Reynolds number, surface roughness, and the shape of the body. Denoting the viscous drag and the pressure drag forces by F_v and F_p respectively one can define two dimensionless drag *coefficients* as $C_{v,p} = \frac{F_{v,p}}{\frac{1}{2}\rho V^2 A}$, where V is the upstream velocity away from the body and A is effective areas presented to the flow (at times, A is defined differently for the two types of drag), and similarly the total drag coefficient C_D as $C_D = \frac{F_D}{\frac{1}{2}\rho V^2 A}$, where $F_D = F_v + F_p$ is the total drag force. Experimentally obtained values for C_D for diverse shapes, Reynolds numbers, and surface roughness parameters can be compared with theoretical estimates or with theoretically predicted trends (such as whether C_D should increase or decrease with Reynolds number in various velocity ranges) in the absence of precise theoretical formulae.

Considering, for the sake of concreteness, an ellipsoidal obstacle in a flow of velocity V along the x-axis of a Cartesian co-ordinate system, with the principal axis measuring a along the x-direction, and with principal axes measuring b, c along the two other directions perpendicular to the flow, one can look at the variation of the drag coefficient $C_D = \frac{F_D}{\frac{1}{2}\rho V^2 bc}$ as a function of the aspect ratio $\frac{a}{b}$, from which one can have a good idea of the shape dependence of the drag.

Of the two extreme cases $\frac{a}{b} \rightarrow 0$ and $\frac{a}{b} \rightarrow \infty$, the former corresponds to a flat plate placed perpendicularly to the flow, for which the viscous drag goes to zero while the pressure drag is entirely due to the momentum imparted to the front surface by the impact of

the flow (the pressure on the rear surface is zero owing to the fact the flow separates at the edge of the plate), which gives $C_D \approx 2$. The other limit ($\frac{a}{b} \rightarrow \infty$) corresponds to a flat plate with its plane parallel to the flow, for which $F_D \approx F_v$, caused by the boundary layer, and C_D has a relatively small value for a fluid of low viscosity.

In between the two extreme cases one obtains the drag coefficient for various different shapes, where one finds that, for streamlined bodies (large aspect ratio) the drag coefficient, dominated by viscous drag, is small for a fluid of low viscosity (or, more precisely, for a laminar flow with a relatively large Reynolds number), while for blunt bodies (small aspect ratio), the drag coefficient, dominated by pressure drag, has a relatively large value. The general nature of variation of the drag coefficient with the aspect ratio is shown in fig. 7-28.

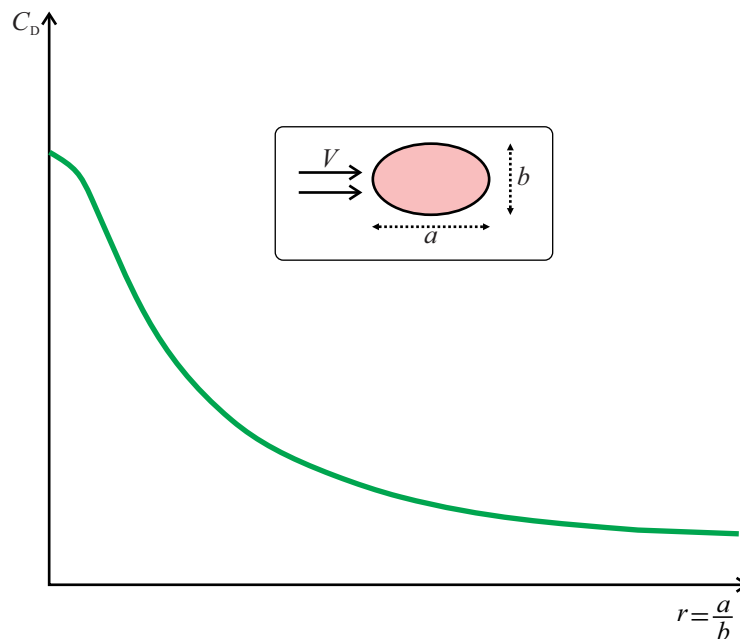


Figure 7-28: Variation (schematic) of drag force exerted by a fluid flowing past an obstacle, with the shape (expressed in terms of aspect ratio); an ellipsoidal obstacle is considered, with axes measuring a, b, c in the direction of the bulk flow and along two perpendicular directions (see inset); the drag coefficient $C_D = \frac{F_D}{\frac{1}{2}\rho V^2 bc}$ is plotted as a function of the aspect ratio $r = \frac{a}{b}$; the limit $r \rightarrow 0$ corresponds to a flat plate placed perpendicular to the flow, for which the viscous component of the drag is nearly zero; the other extreme, $r \rightarrow \infty$ corresponds to a flat plate placed parallel to the flow, for which the pressure drag is negligible; in between these two limits, the drag has a viscous and a pressure component; the plot corresponds to a fixed (and relatively large) value of the Reynolds number, where the flow is streamlined.

The variation of the drag coefficient (C_D) with Reynolds number (\mathcal{R}) for an obstacle of given shape is depicted schematically in fig. 7-29 in the case of a spherical body ($b = c = 2R$; the effective area is taken to be $A = \pi R^2$). For low values of \mathcal{R} (region A of graph) the drag is dominated by viscous resistance (the boundary layer is not well demarcated from the bulk flow) and is given by the Stokes formula of eq. (7-35). The configuration of stream lines in such a flow is shown in fig. 7-30(A), where there is only a small degree of asymmetry between the front and rear surfaces of the sphere. For larger values of the Reynolds number, the viscous drag decreases relative to the pressure drag and the overall drag coefficient decreases slowly (region B of graph in fig. 7-29), being dominated by pressure drag arising due to boundary layer separation. The configuration of stream lines for this region of the graph is shown in fig. 7-30(B) and will be briefly discussed below.

The region C of the graph in fig. 7-29 corresponds to an almost constant value of the drag coefficient, for which the configuration of stream lines is shown in fig. 7-30(C), where the boundary layer is laminar, with a wide turbulent wake. Finally, the region D in fig. 7-29 corresponds to an abrupt dip in the value of the drag coefficient, followed by a constant value, independent of \mathcal{R} , where the flow within the boundary layer as also in the wake is of the nature of a fully developed turbulence, the configuration of stream lines being as in fig. 7-30(D), with a relatively narrow wake behind the rear surface of the sphere.

The configurations of stream lines shown in fig. 7-30(A),(B),(C),(D) correspond to increasing deviations from the configuration for the potential flow past a sphere (refer to fig. 7-18 depicting a section of the potential flow around a circular cylinder; in the case of the sphere, the stream lines look similar in a section through the center of the sphere). In fig (A), the deviation is not marked since the boundary layer is not well demarcated from the bulk flow, and the symmetry between the front and rear surfaces is retained to a considerable degree. For a higher value of the Reynolds number corresponding, for a given flow velocity and a given spherical body, to a low viscosity, the boundary layer is demarcated from the bulk flow, where one observes the boundary layer separation (as at the point E in fig. 7-27) and the backflow in the wake beyond the separation point.

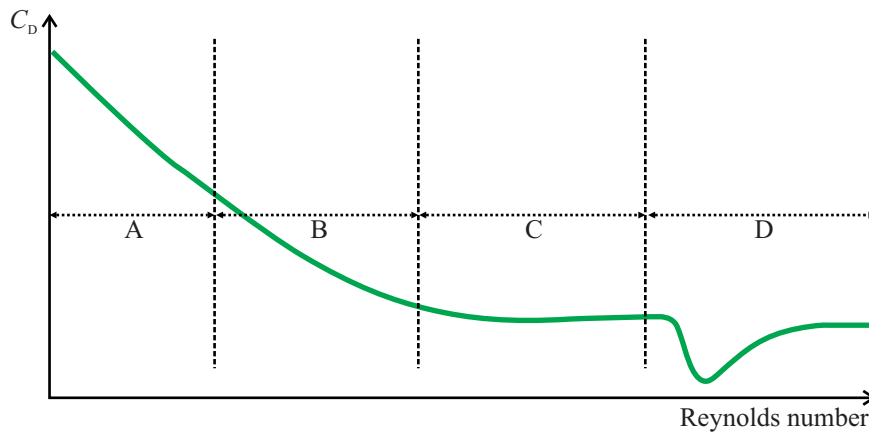


Figure 7-29: Variation (schematic) of drag coefficient exerted by a fluid flowing past a spherical obstacle, with the Reynolds number; the drag coefficient $C_D = \frac{F_D}{\frac{1}{2} \pi \rho V^2 R^2}$ is plotted as a function of the Reynolds number (R); in region A of the graph, which corresponds to the configuration of stream lines shown in fig. 7-30(A), the drag coefficient $C_D (\approx C_v)$ is close to the value given by (7-36a); region B of graph corresponds to the configuration of stream lines shown in figures 7-27 (this figure, though, is for a cylindrical obstacle) and 7-30(B); the drag coefficient shows little variation in region C of graph, which corresponds to a turbulent wake (see fig. 7-30(C)), though the boundary layer flow remains laminar; finally, region D of the graph shows a dip, followed by an almost constant value of the drag coefficient, where the boundary flow becomes turbulent, with a relatively narrow but turbulent wake (see fig. 7-30(D)).

The flow within the wake depicted in figures 7-27 and 7-30(B) is of the laminar type, with the latter corresponding to a higher value of the Reynolds number, where one observes a series of *eddies* within the wake at increasing distances from the obstacle and where, moreover, the *top-bottom symmetry is lost*. This refers to the symmetry between the two sides of a plane, perpendicular to the plane of the figure, containing the line XOX' through the center. One observes alternating eddies on the two sides, with opposite directions of rotation in successive eddies (one each on the top and bottom sides; the terms 'top' and 'bottom' are commonly used to distinguish between the two sides). Such a structure of the wake is referred to as the *Von Karman vortex wake*, in which the flow remains laminar, but is no longer a steady one, being of an *oscillatory* nature.

As the Reynolds number is made to increase beyond the von Karman regime, the flow in the wake assumes a turbulent character, involving a randomly fluctuating series of eddies at various different length scales. The width of the wake is relatively large, and the drag coefficient does not change appreciably with the Reynolds number, corresponding

to the region C in fig. 7-29, as mentioned earlier. The boundary layer flow remains a laminar one, with some admixture of regular flow in the turbulent wake.

The region D in fig. 7-29 begins with a marked dip in the drag coefficient, where the flow is converted to a fully turbulent one. Thereafter, the drag coefficient remains almost independent of the Reynolds number, with fully turbulent flow in the boundary layer and in the wake (fig. 7-30(D)).

Problem 7-9

The Reynolds number for the flow of water past a metal sphere of radius $R = 5.0 \times 10^{-3}\text{m}$ is $\mathcal{R} = 20.0$. If the coefficient of viscosity of water be $8.0 \times 10^{-4}\text{Pa}\cdot\text{s}$, find the velocity of flow and the force of viscous drag on the sphere. If the total drag force is 1.6 times the viscous drag, find the coefficient of pressure drag.

Answer to Problem 7-9

HINT: Using formula (7-36b), one obtains $V = \frac{\eta c a l R}{\rho D}$, where $D = 2R$. Substituting given values of the parameters and $\rho = 10^3\text{kg}\cdot\text{m}^{-3}$, one obtains $V = 4.0 \times 10^{-3}\text{m}\cdot\text{s}^{-1}$. Making use of the formula for viscous drag (eq. (7-35)), the force of viscous drag is found to be $F_v = 3.0 \times 10^{-7}\text{N}$. The total drag force being 1.6 times the viscous drag, one obtains the force of pressure drag as $F_p = 1.8 \times 10^{-7}\text{N}$, and the pressure drag coefficient works out to $C_p = \frac{F_p}{\frac{1}{2}\rho V^2(\pi R^2)} = 0.29$ (approx).

7.5.9 Stability of fluid flow: vorticity and turbulence

The origin and nature of turbulence is a deep and complex question. Generally speaking, turbulence occurs at large values of the *Reynolds number* characterizing a flow, where the Reynolds number is defined as

$$\mathcal{R} = \frac{vL}{\nu}. \quad (7-37)$$

Here v stands for some appropriately chosen velocity and L for some appropriately chosen length characterizing the flow, and $\nu = \frac{\eta}{\rho}$ is the kinematic viscosity of the fluid

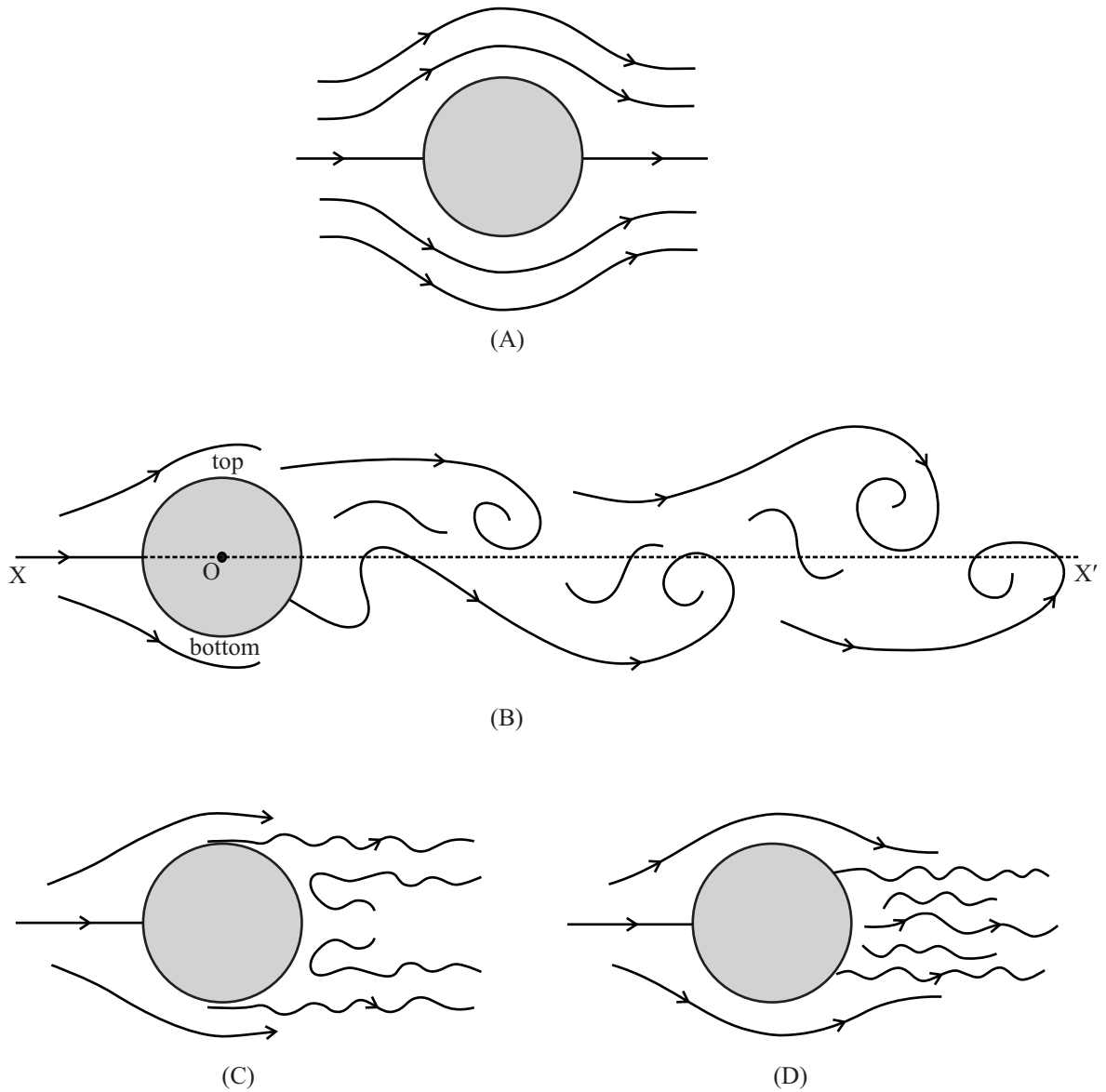


Figure 7-30: Flow past a spherical obstacle for successively increasing values of Reynolds number; (A) configuration of stream lines for low Reynolds number, where there is little asymmetry between the upstream and downstream flows and drag coefficient (dominated by the viscous drag) varies as in (7-36a); boundary layer is not well demarcated from bulk flow; (B) boundary layer demarcated from the bulk flow, where there occurs boundary layer separation as depicted in fig. 7-27; the drag coefficient, with a considerable pressure drag component, decreases more slowly with increasing R than implied by (7-36a); the figure shows the von Karman wake, with alternating eddies shed from the top and bottom halves of the obstacle, and an oscillatory flow; (C) turbulent wake along with predominantly laminar boundary layer flow; (D) fully developed turbulence, with vortices of various length scales dominating the flow; the wake is narrower compared to that in (C).

under consideration. Thus, in the case of a flow past a spherical obstacle, with a uniform upstream velocity, one chooses v to be the upstream velocity V , and L to be the diameter D of the sphere, so as to arrive at the formula (7-36b). Similarly, in the case of Poiseuille's flow through a tube, v can be chosen to be the average velocity of flow and L the diameter of the flow tube (refer to sec. 7.5.6). Finally, in the case of flow past a long flat plate, v is chosen as the upstream velocity (assumed uniform), while L varies from point to point, being the distance from the leading edge of the plate (refer to formula (7-31)).

If the flow occurs under the influence of an external force field or if thermal conduction occurs through the fluid under consideration, then a number of additional dimensionless constants are needed to characterize the flow.

On the face of it, any and every flow of an ideal fluid, for which the coefficient of viscosity (η) or, equivalently, the kinematic viscosity (ν) is zero, is to be of the turbulent type since such a flow is characterized by an infinitely large Reynolds number. This, however, is *not* the case.

The equation of motion of a fluid, whether ideal or not, is, in principle, a *non-linear* one where an exact or an approximate solution arrived at from theoretical considerations may have little relevance from a practical point of view, because such a solution may be *unstable* against small disturbances that can and do occur all the while. For instance, the potential flow of an ideal fluid, introduced in sec. 7.3.7 is, in principle, unstable against a large variety of disturbances (or *perturbations*, as they are referred to). Thus, considering a very small perturbation involving a *vorticity* (see sec. 7.3.3), on a potential flow, the former continues to be present in the flow which, in some sense, continues to differ more and more from the potential flow one started with. Indeed, there is no damping mechanism for the removal of perturbations in the flow of an ideal fluid while, on the other hand, such perturbations can involve a number of modes that can get *amplified* within the flow.

In spite of this, however, potential flows or more general flows of an ideal fluid can serve

as useful approximations to real flows since a real fluid possesses a non-zero *viscosity* that provides a damping mechanism whereby an assumed flow, having idealized features, can be a close approximation to an actual flow that is a *stable* one. This shows that a flow with a large value of \mathcal{R} can be, in principle, qualitatively different from one for which \mathcal{R} is *infinitely* large while, on the other hand idealized flows can have a certain measure of relevance in the description of real ones. Indeed, viscosity plays a significant but subtle role in the phenomenon of turbulence.

the Navier-Stokes equation describing the motion of a real fluid is a nonlinear partial differential equation satisfied by the velocity field, involving ν as a coefficient multiplying the *highest order* partial derivatives of velocity, as a result of which the limit \mathcal{R} turns out to be a *singular* one, which is at the root of the fact that a flow with $\nu = 0$ may differ qualitatively from ones with small but non-zero values of ν . This is also one reason why it requires quite non-trivial considerations to adequately explain and describe turbulent flows of real fluids.

As mentioned earlier, a turbulent flow is one, where the fluid velocity and pressure at any point in the flow (or in a certain region within the flow) varies irregularly, as opposed to a laminar flow where the fluid velocity at any point is a regular function of time, being a constant in the special case of a steady flow. For a turbulent flow, one refers to mean values of velocity and pressure at any given point in order to describe the flow. Under a given set of circumstances, a laminar flow becomes unstable as the Reynolds number is made to increase where, moreover, there may appear a *succession* of instabilities till there emerges a fully developed turbulence. In other words, *viscosity* and *stability of flow* constitute two key concepts in the explanation of turbulence.

The third characteristic feature of fluid flow that emerges as one of crucial significance in the understanding and explanation of turbulence is *vorticity*.

Indeed, turbulence can be defined as a flow involving a cascade of vorticities of varying length scales where small scale vorticities are fed by large scale ones so as produce a chaotic spatial and temporal flow pattern, involving viscous dissipation at the lower end

of the spectrum of length scales. This leads to a dissipation of energy associated with large scale vortices into energy of microscopic molecular motions.

As for the large scale vortices, viscosity is only one of the numerous factors that are responsible for their continual generation and break-up. Generally speaking, large scale features of vortex dynamics specific to turbulence, do not depend significantly on viscosity.

However, the *creation* of large scale vortices is principally dependent on viscosity. Eddies (whorling masses of a fluid, generally possessing non-zero vorticity) are created near boundary surfaces of solids with which a viscous fluid is in contact, and move along with the flow of the latter (this is referred to as 'advection'), thereby spilling over into relatively distant regions. While the eddies (or 'vortices') are nothing but velocity fields with a special structure, these have an identity of their own and can be treated, in a sense, as physical objects moving about within the flow (this comes about in virtue of the principle of conservation of angular momentum). An eddy creates around itself a velocity field of a certain description (analogous to the magnetic field around a current) while, on the other hand, the velocity field in the region within which an eddy has been injected influences and modifies the eddy structure.

In other words, the non-linearity of the equation describing fluid flow can be alternatively looked at as an *interaction* between eddies and velocity fields where the latter can be taken to be of relatively simple forms such as ones describing irrotational flows. This mutual interaction, in turn, leads to an interaction *among the eddies*. The resulting *vorticity dynamics* is of crucial significance in turbulence.

Analogous to the stream lines describing a velocity field, a vortex field (vorticity distribution in space; recall that the vorticity is nothing but the curl of the velocity field at a point) can be described in terms of *vortex lines*, where a bunch of adjacent vortex lines form a vortex *tube*. A convenient description of the flow of a fluid of a low viscosity is one in terms of the stretching, twisting, and advective motion of a large number of vortex tubes swirling about in the flow. The vortex tubes start and end on solid boundary

surfaces or else have a ring-like structure, each tube having a 'strength' (a measure of vorticity) of its own that can increase (by stretching) or decrease (by spreading) while other modes of evolution such as amalgamation, diffusion, and the creation of one vortex pattern from another are also of likely occurrence. While vortices are identifiable structures within a flow, they have a certain turnover time in which they break up and produce new vortices, thereby creating a complex spatial and temporal pattern of evolving tangles of vorticities.

A considerable insight into the variety and richness of vortex dynamics can be gained by considering vortices in *two dimensional* flows, where it is found that, even in such relatively simple flows, the dynamics evolves into a *chaotic* one, resembling the irregular dynamics of turbulent motion. However, even the chaotic dynamics of a two dimensional vortical flow does not capture the essence of turbulence since the former does not incorporate the mechanism of *energy dissipation* characterizing the latter.

Vortices in three dimensions have an extra degree of freedom whereby these can undergo the process of *vortex stretching* that results in an increase of their strength (once again a consequence of the principle of conservation of angular momentum). A measure of the strength of a vortex is the *enstrophy* associated with it, where the enstrophy is defined as $\frac{1}{2}\omega^2$ ($\vec{\omega} \equiv \text{curl } \mathbf{v}$) (with an appropriate averaging). It is the enstrophy that determines the rate of energy dissipation in the flow, and ultimately connects the energy dissipation in turbulence with the phenomenon of vortex stretching in three dimensions. The process of stretching generates vortex motion in ever finer length scales and ends up in the production of very large, effectively *infinite*, velocity gradients. Such large velocity gradients, in turn, produce viscous dissipation at small lengths which, however, is of an anomalous nature in that the dissipation is effectively independent of viscosity (recall from sections 7.5.8.2 and 7.5.8.3 how the viscous drag coefficient becomes independent of the Reynolds number for fully developed turbulence).

Let us imagine a mass of uniformly moving fluid which encounters one or more obstacles in its path whereafter it is again allowed to flow unhindered through a wide tunnel. It will then be found that the flow undergoes a number of transitions as it moves further

and further away from the obstacles. Initially, the flow will consist of only large scale vortices whose size is determined by the size of the obstacles and of the tunnel. But then the vortex dynamics takes over, where vortices are formed at smaller and smaller length scales and one observes a phase of fully developed turbulence, with the spectrum of length scales extending down to the so-called *Kolmogorov scale*. In this fully developed turbulence, energy is distributed mainly among the large scale vortices while enstrophy is concentrated in the ones at small length scales. Once all the length scales are generated, there begins a phase of *decay* of turbulence when the energy of the small scale vortices gets lost into random molecular motion by viscous dissipation.

It was A. N. Kolmogorov who famously gave an insightful theory of turbulence based on length scales. Kolmogorov's theory complements the theory, outlined by L. F. Richardson, of energy transfer across length scales from larger to smaller eddies. However, in spite of subsequent improvements and developments of the theory initiated by these pioneers, involving statistical descriptions of the turbulent velocity field, a satisfactory theory of turbulence still remains elusive. It is perhaps in the very nature of turbulence that this should be so.

7.6 Surface energy and surface tension

7.6.1 Introduction

The boundary surface of a body affects, to some extent, its behavior and its characteristic features. Commonly, however, most of the molecules in a body are located within the bulk of the material making up the body and only a small fraction of the molecules are located near the boundary surface, as a result of which the effect of the surface on the behavior of the entire body is not pronounced. The effect is nevertheless there, and becomes appreciable in certain situations when the interfaces between materials are found to be significant in determining the behavior of macroscopic systems.

The boundary surface of a body usually separates two different media, while in certain situations more than two media may also be involved. The effect of the boundary surface

comes about through interactions involving molecules on both the sides of this surface. Strictly speaking, in referring to the surface properties of a liquid, it is to be assumed that the medium on the other side of the boundary surface is the saturated vapor of that liquid. The saturated vapor possesses the special property that the liquid can remain in equilibrium in contact with its own saturated vapor (see section 8.22) and thermodynamic parameters can be unambiguously defined for the liquid-vapor system.

In practice, however, one can consider the liquid surface exposed to air or some other gaseous medium, since ordinarily the energy of a liquid molecule located near its surface does not change appreciably if the medium on the other side of the surface is made to change from the saturated vapor to air or a gaseous medium. Thus, the surface tension coefficient (see below) of the liquid that governs its surface properties does not, under ordinary circumstances, depend appreciably on whether the medium in question is any chosen gas or vapor, and it will not be necessary in the following to explicitly refer to it. Strictly speaking, though, a liquid with its surface in contact with an arbitrarily chosen gas cannot be in equilibrium, and only the liquid in contact with its saturated vapor constitutes a system in thermodynamic equilibrium.

Situations may, however, arise in which the surface tension with reference to the saturated vapor does differ appreciably from the one defined with reference to some other gaseous medium. For instance, at the *critical temperature* of the liquid (see sec. 8.22.5), the surface tension with reference to the saturated vapor vanishes while that with reference to air or vacuum does not.

7.6.2 Surface energy and surface tension: thermodynamic considerations

The environment of a molecule near the boundary surface of a liquid is quite different compared to that of a molecule in the bulk of the liquid, away from the surface. For instance, a molecule well in the interior of the volume occupied by the liquid is surrounded on all sides by molecules of the same kind while one near the surface finds neighboring molecules belonging to the liquid medium only on one side, the molecules

on the other side being in a different state of organization since they belong to a different medium (fig. 7-31).

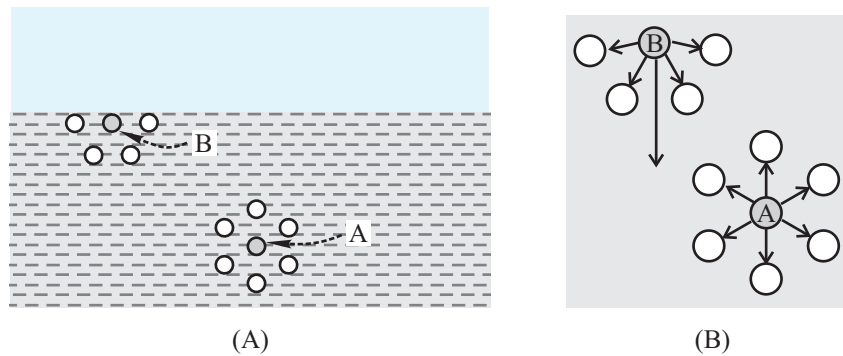


Figure 7-31: A molecule A in the interior of a liquid compared to a molecule B near the surface: (A) the two molecules have different environments, (B) while molecule A is pulled from all sides equally by other liquid molecules, molecule B experiences pull exerted by other liquid molecules from one side only.

In general, a molecule experiences an unbalanced force toward the interior of the liquid as it is made to move from the interior towards the boundary surface, as a result of which some amount of energy is to be given to it in order to bring it from the interior to the surface (see fig. 7-31(B)). In other words, molecules making up the surface of the liquid possess a higher energy compared to molecules in the interior, the latter making up the bulk of the liquid. To put it another way, it requires a supply of energy from some source to increase the surface area of the liquid by way of more molecules being transferred from the bulk to the surface.

Suppose the surface area of a certain mass of liquid is increased by an amount δA by an *adiabatic* process, i.e., a process where *no heat exchange* takes place between the liquid surface and other bodies making up its environment (see chapter 8 for an introduction to the concept of heat exchange and adiabatic processes), but one where energy can be given to or taken away from it in the form of *work*, keeping, however, the volume and the number of moles unchanged. Suppose that the amount of work performed in the process is δW . Then the ratio $\frac{\delta W}{\delta A}$, i.e., the work performed per unit increase in the surface area in an adiabatic process, is termed the *surface tension coefficient*, or simply,

the surface tension, of the liquid:

$$\gamma = \left[\frac{\delta W}{\delta A} \right]_{\text{adiabatic}}. \quad (7-38)$$

The unit of surface tension is $\text{N} \cdot \text{m}^{-1}$. The surface tension of a liquid depends on its temperature. For instance, the surface tension of water at 293 K is nearly $0.073 \text{ N} \cdot \text{m}^{-1}$.

1. The surface tension is a characteristic of the surface looked upon as a thermodynamic system, with the liquid on one side and some medium on the other side. If the medium on the other side is the saturated vapor of the liquid or vacuum, then the surface tension reduces to a characteristic of the liquid alone. Considering a surface with two different materials on either side of it, the surface energy per unit area defined as above is referred to as the *interfacial* surface tension for the materials under consideration.
2. The concept of a sharply defined two dimensional surface separating two media is an idealization and was introduced by J. W. Gibbs in his pioneering work on the thermodynamics of surfaces. Gibbs' definition of a surface is a subtle and conceptual one, which I will not enter into. Instead, I will assume in the following that the surface is well defined and has a physical existence.

Denoting the surface tension of a liquid by γ , one then concludes that the work required to expand the surface area by δA in an adiabatic process is $\gamma \delta A$, where it is assumed that the volume and the mole number are not changed. Conversely, in a contraction of surface area, the surface is capable of delivering work to an external system. More usually though, one is required to consider the expansion or contraction of a surface in an *isothermal* process where, in addition to work being done on or by the surface, there occurs heat exchange between the surface and the surrounding system(s), with the temperature remaining unchanged (see, once again, chapter 8 for an introduction to some of the concepts mentioned here). In such a situation, the quantity $\gamma \delta A$ equals the change in the *thermodynamic free energy* of the system, where the volume and the mole number are again assumed to remain unchanged. However, I will not enter here into a detailed thermodynamic consideration relating to the surface of a liquid.

1. The concepts of adiabatic and isothermal processes, heat exchange, work, and a number of other related concepts, are of great relevance in the science of *thermodynamics*. One fundamental principle in thermodynamics, known as the *second law of thermodynamics*, tells us of a natural tendency of systems to exchange heat and work with other systems in their environment in such a manner that a certain quantity, known as *entropy* of all the systems taken together, goes on increasing. Based on this principle, thermodynamics provides us with a concrete set of criteria that can be made use of to describe or to predict the way a system exchanges work and heat under a given set of conditions.

For instance, suppose a system exchanges energy with a thermal reservoir in such a manner that its initial and final temperatures are the same as that of the reservoir, which may be assumed to remain unchanged in the process. Then, assuming that no work is performed on the reservoir, the *free energy* of the system (another thermodynamic concept) must *decrease* in the process because that is the only way the total entropy of the system and the thermal bath can increase.

The reason I am telling you all this is that the behavior of the liquid surface is ultimately tied up with these thermodynamic principles. And it is thermodynamics that also tells us that the free energy associated with a liquid surface is, simply, γA , where A is the area of the surface and where, for a given liquid, the surface tension γ is a function of the temperature alone. This leads us to the conclusion that I am now going to state below (sec. 7.6.3).

2. The value of the surface tension of a liquid depends on the force of interaction between its molecules. In general, the intermolecular forces are of electromagnetic origin and can mostly be described as being due to the *van der Waals* interaction. The van der Waals interaction involves forces due to the electrical dipole moments (see chapter 11) of molecules and may arise between two molecules having dipole moments of their own, between a molecule having a dipole moment and one having no dipole moment of its own, or between two molecules none of which possesses a dipole moment of its own. In the last instance the interaction force (commonly referred to as the London dispersion force) arises due to a quantum mechanical effect (see chapter 16 for an introduction to the basic ideas in quantum theory) involving *fluctuating* dipole moments in the two interacting molecules, the average

dipole moment of each molecule being zero. In the case of water, the polar nature of the water molecules plays a special role in determining its surface tension, which has a high value compared to most other liquids.

7.6.3 The tendency of a liquid surface to shrink

If, under a given set of conditions, a liquid surface is made to remain in thermal contact with its surroundings at a given temperature, and if no work is performed on the liquid through a change in its volume or mole number, then the area of the surface tends to decrease till an equilibrium is reached and the area of the surface assumes the *minimum* value under the given conditions. For instance, if a liquid drop of given volume is in thermal contact with the atmosphere and if no external force acts on the drop then the shape of the drop will change spontaneously in such a way that its surface area will decrease and, in the end, it will attain a shape with the *minimum* possible surface area which, for a drop with a given volume, is that of a *sphere*. The bulk of the drop and the atmosphere here act as a thermal reservoir exchanging heat with the liquid surface.

As another example, consider a thin film of liquid stretched between a thin rigid wire frame, the latter being confined to a plane, as in fig. 7-32(A). Here the surface is in thermal contact with the atmosphere which we assume to be at a constant temperature. Then, supposing that the shape of the surface is as in fig. 7-32(A) to start with, with part of the surface lying outside the plane, it will have the tendency to undergo a spontaneous change of shape so as to finally become a plane surface, lying in the plane of the wire frame as in fig. 7-32(B). In practice, however, it may not be possible to observe the flattening of the film since the latter is likely to lose its integrity before it attains the configuration of minimum surface area, because of the lack of the excess pressure necessary to support it (see sec. 7.6.6). On the other hand, if the film is flat to start with, then it will be found to remain in equilibrium, since the flat configuration is the one of smallest area consistent with the fixed perimeter defined by the frame.

This tendency of a liquid surface to shrink to a minimum area is a consequence of the fact that a liquid molecule located within the bulk of the liquid possesses a lower energy

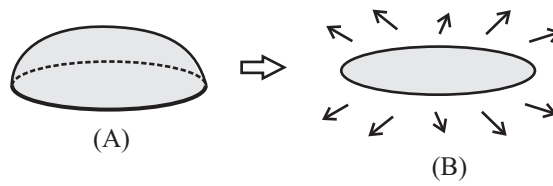


Figure 7-32: Natural tendency of a liquid film to shrink in surface area; the film, originally as in (A), tends to shrink so as to lie in the plane of the frame as in (B); the frame exerts a pull on the film as shown by the direction of the arrows in (B).

compared to one located near the surface. Referring to the above example of the film stretched between the wire frame, the surface area would decrease further if the frame were made of moveable parts instead of being rigid (fig. 7-33). The fact that a rigid frame prevents the film from shrinking in area means that the frame must be exerting a stretching force or lateral pull, in the direction of the arrows shown in fig. 7-32 (B), on the surface. In other words, the surface acts as an elastic membrane, tending to shrink in area unless a stretching force prevents it from shrinking.

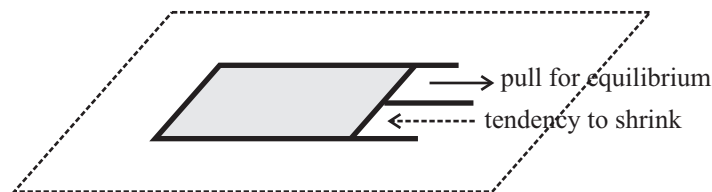


Figure 7-33: Tendency of a liquid film to shrink in area; a stretching force exerted on the movable part of the wire frame keeps the film in equilibrium.

7.6.4 Surface tension as lateral force

Based on the above considerations, the surface tension coefficient γ can be alternatively defined in terms of the stretching force required to maintain the surface in equilibrium.

Imagine, then, a portion of a liquid surface bounded by a closed curve as in fig. 7-34(A), where forces depicted by arrowheads are applied to this portion of the surface so as to keep it in equilibrium. If the portion under consideration is part of a bigger surface then these forces would be caused by interactions with surrounding liquid molecules while, for a situation such the one depicted in fig. 7-33 the forces arise due to the pull exerted

by other bodies like the wire frame.

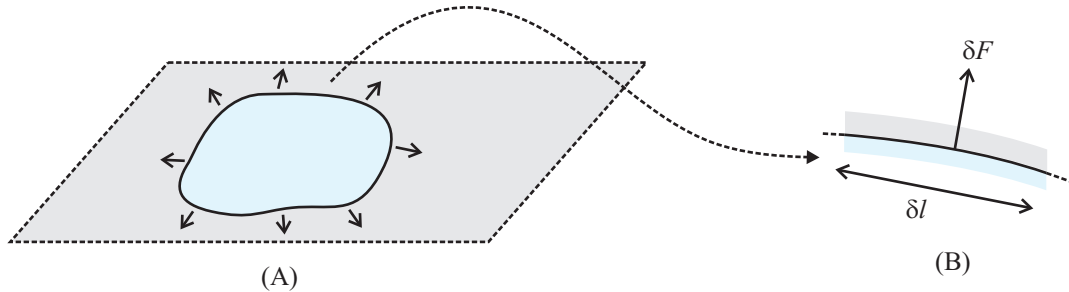


Figure 7-34: (A) Lateral pull required to keep a part of a liquid surface, bounded by a closed curve, in equilibrium; (B) lateral force on a small element on length δl on the boundary, in terms of which one can define the surface tension coefficient γ .

Looking at a small element of length δl on the boundary of the closed area under consideration (fig. 7-34(B)), let δF be the pull, directed normally to the length element and lying in a plane tangential to the liquid surface, along with similar pulls elsewhere on the closed boundary curve, that keeps the closed area in equilibrium. Then the surface tension coefficient (γ) is simply the pull per unit length:

$$\gamma = \frac{\delta F}{\delta l}, \quad (7-39)$$

which is consistent with the definition (7-38) (check this out).

7.6.5 Angle of contact

Considerations relating to surface tension of a liquid and associated phenomena belong to the broad area in physics referred to as *surface physics*. Surface physics includes a wide range of features and phenomena pertaining to surfaces of materials in contact, and requires a broad-based theoretical approach to describe the properties of such surfaces. Such a broad-based approach tells us, for instance, that the property of surface tension is not peculiar to liquids alone and is relevant, for instance, for a solid-gas interface as well.

It may be mentioned in particular that surface tension is related to the property of

adhesion between dissimilar materials having a common interface (in addition to being related to the property of *cohesion* as well), where adhesion, which involves a tendency of two distinct materials to stick together at their interface, may be manifested through diverse features and phenomena. It determines the surface tension (γ) of the interface where γ is once again defined as $\frac{\delta W}{\delta A}|_{\text{adiabatic}}$, i.e., the work necessary per unit increase in the interface area in an adiabatic process, the latter being a process involving no exchange of heat.

Consider now a situation like the one shown in fig. 7-35 depicting a drop of mercury resting on a horizontal glass plate, where *three* interfaces are involved - the mercury-glass surface, the mercury-air surface, and the glass-air surface. All the three interfaces come together at the *line of contact*, which is the line marking the boundary of the bottom surface of the mercury drop.

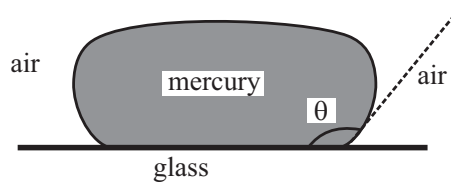


Figure 7-35: A pellet of mercury on glass, showing an obtuse angle of contact (θ).

The equilibrium configuration of the entire system made up of the three interfaces is determined by the surface tensions arising from the forces of cohesion relating to the three materials under consideration, and adhesion between the three pairs of materials, as also on the system of external forces acting on the system, which in this case is made up of the gravitational force on the mercury drop. In particular, the *angle of contact* (θ) between mercury and glass is determined by the three surface tensions (γ_1 , γ_2 , and γ_3 , being respectively the surface tensions pertaining to the mercury-air, mercury-glass, and glass-air interfaces) alone. Note from the figure that the angle of contact is defined as the angle, measured within the liquid, between the liquid-gas (i.e., mercury-air in the present example) surface and the solid-liquid (glass-mercury) surface, and is an *obtuse* angle in the case of mercury-glass contact (with air as the third medium). In this

context, one commonly refers to the following formula relating γ_1 , γ_2 , and γ_3 with the angle of contact θ under equilibrium conditions, known as the *Young formula*,

$$\gamma_1 \cos \theta + \gamma_2 - \gamma_3 = 0. \quad (7-40)$$

Fig. 7-36 depicts the surface of water in contact with a vertical glass plate where, once again, air is the third medium in contact with water and glass. One finds that the angle of contact (θ) is an *acute* angle in this case. The proper interpretation of the Young formula, however, is not simple since one has to include here a number of factors such as the *adsorption* of the three media into one another. Additionally, for small drops of a liquid on a solid surface, one has to take into account the *line tension* along the common line of separation, since this line acts as a stretched string, along which a force of tension acts at every point.

It is the angle of contact between a liquid and a solid that acts as a factor determining whether or not the liquid *wets* the surface of the solid. The angle of contact between mercury and glass being an obtuse one, mercury does not wet the surface of glass and collects on a glass plate in the form of pellets or pearls. On the other hand, water wets the surface of glass because of the fact that the angle of contact is an acute one. Thus, when water is poured on a horizontal glass plate, it spreads out to form a thin film on the plate, as in fig. 7-37 (strictly speaking, the equilibrium value of the angle of contact is zero for complete wetting). Ordinarily, the angle of contact of a liquid with a given solid surface decreases with decreasing surface tension of the liquid or, more precisely, of the liquid-gas (commonly a mixture of air and the liquid vapor) interface.

The angle of contact depends strongly on the surface conditions of the materials in contact. Thus the spreading of water on a glass plate is affected to a marked degree by the glass plate being contaminated with greasy or waxy materials. Certain materials, when added to liquids, alter remarkably their surface properties, including their angles of contact with other surfaces and their ability to wet and spread over these surfaces. These

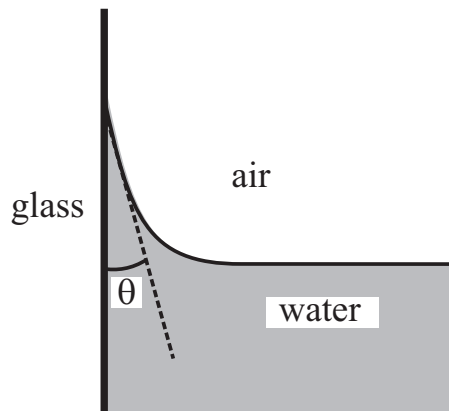


Figure 7-36: Water in contact with a vertical glass plate showing acute angle of contact (θ).



Figure 7-37: Water spreading on a horizontal glass plate as a thin film; successive stages of spreading are shown in (A), (B), (C).

materials, termed *surfactants* are widely used for household and industrial purposes (see sec. 7.6.8).

The theory describing various aspects of wetting of surfaces by liquids involves complex considerations in surface physics. The phenomenon of wetting itself is an intricate one. For instance, even for a clean surface, it depends on the degree of roughness and homogeneity of the surface under consideration, and may involve states that are described as *metastable* ones from the thermodynamic point of view. The process of wetting is characterized by phenomena such as *hysteresis*, where the contact angle of a spreading drop of liquid differs from that of a retracting drop.

The current scientific understanding of phenomena relating to interfaces is not as developed as the theory underlying the atomic and molecular constitution of matter on the one hand, and of the understanding of bulk properties of materials on the other. Interfaces and *colloids* are of intermediate dimensions compared to the two extreme regimes of aggregation of matter - atoms and molecules at one end, and bulk materials at the other. This is why surface and colloid physics, having had only a moderate growth in the past, has been a rapidly developing field of study during recent decades,

spurred by a wide range of potential and actual applications. Numerous features of surface phenomena are still not understood adequately, one reason for this being the fact that simplified or idealized models are often of little value in describing real interfaces. Put differently, real interfaces are inherently complex objects of study.

The theory of surface phenomena for *solids* (in contact with media of various other descriptions) is of great relevance. In the case of a solid, the concept of surface tension is to be generalized to one of a surface *stress* where the surface force acting through an imagined line on the surface does not necessarily act in a direction perpendicular to the line.

7.6.6 Pressure difference across a curved liquid surface

Since the surface of a liquid resembles a stretched membrane, a pressure difference is necessary to maintain a curvature of such a surface like, for instance, the pressure necessary to blow a balloon. Since the surface of a liquid in contact with a solid surface often assumes the form of a concave or convex meniscus (see sec. 7.6.7 and fig. 7-36), such pressure differences are commonly seen to arise across liquid surfaces. The analogy with a blown balloon is particularly transparent in the case of a liquid film or a bubble.

Consider a curved liquid surface as in fig. 7-38(A) where the medium on the convex side of the surface is air while that on the concave side is the liquid under consideration. Assuming the surface to be a spherical meniscus of radius r , the pressures p_A and p_B at the two points A and B shown in the figure are related as

$$p_A - p_B = \frac{2\gamma}{r}, \quad (7-41)$$

where γ stands for the surface tension of the liquid under consideration. In other words, there has to be an excess pressure on the concave side of the surface as compared to the convex side so that the curved surface can remain in equilibrium. Fig. 7-38(B) shows a curved surface with air on the concave side and a liquid on the convex side with A and B being points in the liquid and in air respectively. In this case the pressures at the two

points are related by

$$p_A - p_B = -\frac{2\gamma}{r}, \quad (7-42)$$

which means, once again, that there has to be an excess pressure on the concave side in order that the curved surface may remain in equilibrium. Eq. (7-42) is commonly referred to as the *Laplace formula* for a curved surface.

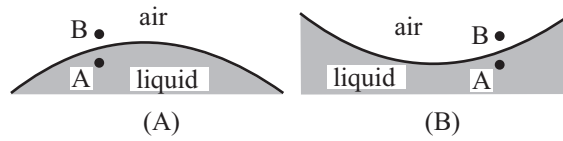


Figure 7-38: Illustrating the pressure difference across a curved liquid surface; (A) liquid on concave side, (B) liquid on convex side; the pressure difference between the points A (in liquid) and B (in air) is given by eq. (7-41) in (A) and (7-42) in (B).

If the curved surface under consideration is not spherical in shape, then one can define two *principal radii of curvature* (say, r_1 and r_2) at any point on it, and the pressure difference across the surface at the point will then be given by

$$\delta p = \gamma \left(\frac{1}{r_1} + \frac{1}{r_2} \right), \quad (7-43)$$

Here δp is the pressure difference between two points A and B ($\delta p = p_A - p_B$), one on each side of the surface under consideration, and either of r_1 and r_2 is taken to be positive (resp. negative) if the corresponding line of curvature is concave (resp. convex) when looked at from the side of A. For a spherical surface the two principal radii of curvature are equal, while for a cylindrical surface, one of the two principal radii is infinitely large.

Fig. 7-39 shows part of a liquid film of spherical shape, with the thickness magnified for the sake of clarity. Considering points A, B, C as shown in the figure, one has

$$p_A - p_B = p_B - p_C = \frac{2\gamma}{r}, \quad (7-44a)$$

i.e.,

$$p_A - p_C = \frac{4\gamma}{r}. \quad (7-44b)$$

Thus, for instance, the excess pressure in the interior of a spherical soap bubble of radius r as compared to its exterior has to be $\frac{4\gamma}{r}$ in order that the bubble may be in equilibrium, where γ stands for the surface tension of the soap solution.

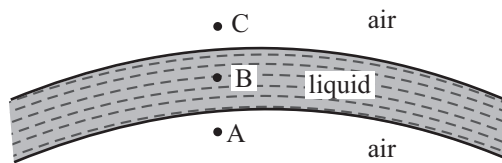


Figure 7-39: Illustrating the pressure difference across a liquid film of spherical shape; a part of the film is shown, with the thickness magnified; A and C are points in air, on the concave and convex sides of the film respectively, while B is a point in the liquid; the pressure difference between A and B, and that between B and C, are given by eq. (7-44a).

Problem 7-10

Air bubbles are formed at two ends of a long narrow tube through which the flow of air, if any, occurs very slowly. If the surface tension of the liquid of which the bubbles are made be $\gamma_1 = 0.030 \text{ N}\cdot\text{m}^{-1}$ and $\gamma_2 = 0.040 \text{ N}\cdot\text{m}^{-1}$, and the radii of the bubbles are respectively $r_1 = 0.01 \text{ m}$ and $r_2 = 0.02 \text{ m}$ respectively, in which direction will air flow through the tube, and what will be the initial pressure difference driving the flow of air? If, at any intermediate stage, the radii of the bubbles be, respectively, r'_1 and r'_2 , then set up an equation relating the two.

Answer to Problem 7-10

Since the flow of air through the tube occurs slowly, the bubbles may be assumed to be in equilibrium, in which case the excess pressure in each bubble can be obtained from formula (7-44b). If P_0 be the atmospheric pressure, then the pressures in the two bubbles are respectively $P_1 = P_0 + \frac{4\gamma_1}{r_1}$ and $P_0 + \frac{4\gamma_2}{r_2}$. Here $\frac{4\gamma_1}{r_1} = 12.0 \text{ N}\cdot\text{m}^{-2}$, and $\frac{4\gamma_2}{r_2} = 8.0 \text{ N}\cdot\text{m}^{-2}$. Thus, air will flow through the narrow tube from the smaller to the larger bubble, driven by an initial pressure difference of $4.0 \text{ N}\cdot\text{m}^{-2}$.

Since the flow through the tube is assumed to be slow one, at each point of time the air in either of the two bubble may be assumed to be in thermal equilibrium (refer to chapter 8 for general background) with the surrounding air at an absolute temperature, say T , in which case its radius r will satisfy $(P_0 + \frac{4\gamma}{r})\frac{4}{3}\pi r^3 = \nu RT$ (refer to formula (8-16); notation as explained there, i.e., ν is the mole number of air in the bubble, and R is the universal gas constant), where γ stands for the relevant value of surface tension. Since the total mole number of air in the two bubbles has to remain constant during the flow, one obtains,

$$(P_0 + \frac{4\gamma_1}{r_1})r_1^3 + (P_0 + \frac{4\gamma_2}{r_2})r_2^3 = (P_0 + \frac{4\gamma_1}{r_1'})r_1'^3 + (P_0 + \frac{4\gamma_2}{r_2'})r_2'^3,$$

which is the required relation.

7.6.7 Capillary rise

Referring to fig. 7-36, one observes that the water in contact with the vertical glass plate rises to some extent along the plate, forming a *meniscus* where this rise can be interpreted as water wetting the plate against the downward pull of gravity. This rise of a liquid in contact with a surface, observed with reference to the liquid level far away from the surface is referred to as *capillary rise* because it is observed commonly for liquids in narrow tubes or pores. The surface in contact with the liquid has to be such that the liquid wets the surface (i.e., the angle of contact has to be an acute one).

Fig. 7-40(A) shows the capillary rise of water kept in a flat vessel with a narrow open-ended glass tube dipped in it, where it is found that the water rises in the tube above the level in the flat vessel, and the top of the water column in the tube forms an almost hemispherical concave meniscus in the tube (menisci are also formed where the water comes in contact with the walls of the flat vessel). Evidently, this phenomenon of capillary rise is possible only for those liquids that wet the surface of the capillary tube or pore. Fig. 7-40(B) depicts a narrow open-ended tube dipped in mercury kept in a flat trough, where it is found that the mercury level inside the tube is depressed compared to that in the trough, with the meniscus inside the tube being a *convex* one. The explanation lies in the fact that the surface tensions of the three surfaces involved

(mercury-glass, glass-air, and mercury-air) are such that the angle of contact between mercury and glass is an obtuse one.

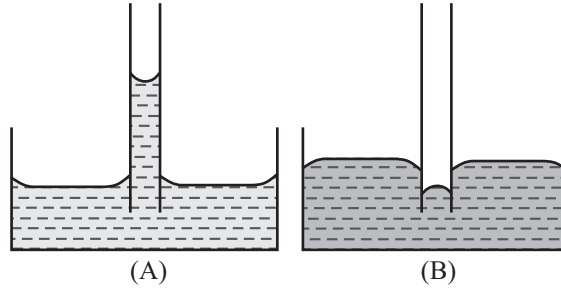


Figure 7-40: (A) Capillary rise of water in a narrow tube dipped in water in a flat trough, (B) depression of mercury level within a narrow tube dipped in mercury in a flat trough; curvatures are magnified for the sake of clarity.

The height to which a liquid rises in a capillary tube can be worked out by considering the surface tension forces brought into play at the free surface of the liquid in the tube. An approximate expression for the height of ascent is

$$h = \frac{2\gamma \cos\theta}{r\rho g}, \quad (7-45)$$

where γ stands for the surface tension of the liquid-air interface, θ is the angle of contact of the liquid with the material of the tube, r is the radius of cross-section of the tube, ρ is the density of the liquid, and g denotes the value of the acceleration due to gravity. As is evident from the formula, the smaller the angle of contact, the larger will the height of ascent be. Another notable feature of the formula is that the height of ascent in the capillary tube increases in inverse proportion to the radius of the tube (*Jurin's law*). Formula (7-45) relates to the capillary rise of a liquid in the equilibrium configuration, while the *dynamical* features of capillary rise require a number of additional considerations for an adequate description, especially for a viscous liquid.

Formulae (7-40), (7-41), (7-42), and (7-45) are all derived by making use of the condition that the free energy of liquid surface in contact with one or more media is to be a minimum under the constraint of constant temperature, volume and mole number(s).

Problem 7-11

A sufficiently long capillary glass tube of radius $r = 5.0 \times 10^{-4}\text{m}$ is dipped into water of surface tension $\gamma = 7.0 \times 10^{-2}\text{N}\cdot\text{m}^{-1}$. Assuming the angle of contact with glass to be 0, obtain the height (h) of capillary rise. If the length of the tube above the free water surface is $h' < h$, explain whether water will flow out from the top of the tube.

Answer to Problem 7-11

HINT: Referring to formula (7-45), and substituting values (in SI units, $\gamma = 7.0 \times 10^{-2}$, $\cos \theta = 1$, $g = 9.8$, $\rho = 1.0 \times 10^3$), one obtains $h = \frac{2\gamma}{r\rho g} = 2.9 \times 10^{-2}\text{m}$ (approx).

If the height of the tube above the free surface of water is $h' < h$, water does not flow out of the tube since, in order to flow out, the curved meniscus will first have to be flattened in which case the angle of contact will be $\frac{\pi}{2}$. Before this can happen, a configuration will be reached in which the meniscus is less curved compared to that in fig. 7-41(A), where the full height of ascent h is achieved, the angle of contact being now θ (see fig. 7-41(B)) such that $h' = \frac{2\gamma \cos \theta}{r\rho g}$. In other words, $\theta = \arccos \frac{h'}{h}$. The smaller the value of h' , the more flattened will the meniscus be.

NOTE: If water could flow out from a tube of height smaller than h , then that could be made use of in devising a *perpetual motion machine*, in contravention of the *first law of thermodynamics*.

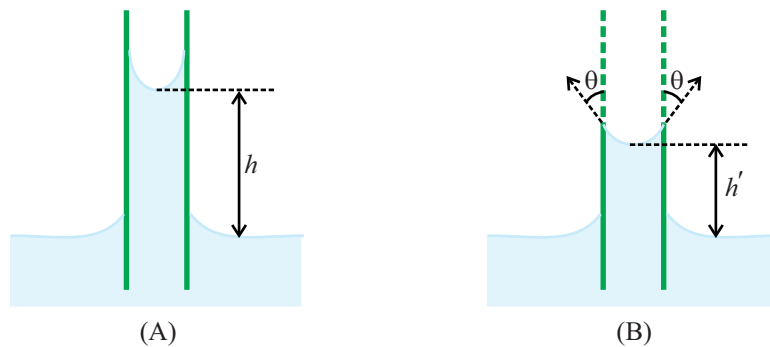


Figure 7-41: (A) Long capillary glass tube dipped in water; water wets the surface of glass and rises up to height h by capillary action, where h is given by (7-45), with $\theta = 0$; (B) a shorter tube of the same radius is dipped, such that the length of tube above the free surface of water is $h' < h$; in this case, the water will rise to a height h' , but will not flow out of the tube (problem 7-11); the meniscus will become flattened in comparison to that in (A), and the angle of contact θ will get changed to a non-zero value, being given by $\theta = \arccos \frac{h'}{h}$; figure (schematic) not to scale.

Incidentally, working out the *shape of the meniscus* within a capillary tube or, more generally, of one formed at the surface of a solid boundary in contact with a liquid, constitutes a non-trivial problem. Strictly speaking, the curvature at all points on the curved surface of the meniscus within a capillary tube need not be constant, i.e., the surface need not be spherical. Accordingly, in formula (7-45), which is an approximate one (in which the weight of a small volume of liquid, bounded by the meniscus surface and a horizontal surface through the lowest point of the meniscus, is ignored), r is to be interpreted as the radius of curvature at the lowest point on the meniscus, and its equality with the radius of the capillary tube can be assumed only as an estimate.

Problem 7-12

Attraction between two plates dipped in liquid.

Two flat rectangular plates of the same material are dipped in a liquid with their planes parallel. Show that they will experience a force of attraction to each other due to the surface tension of the liquid, regardless of the angle of contact with the material of the plates (see fig. 7-42), and make an estimate of this force.

Answer to Problem 7-12

HINT: Referring to fig. 7-42, there are two cases to consider, namely (A) an acute angle of contact between the liquid and the plate, for which the liquid rises up between the two plates (up to level M), and (B) an obtuse angle of contact, for which the liquid level is depressed within the two plates (down to M). In (A), considering any one of the plates (say, the one marked A), the forces on it due to air pressure from the two sides above the level M cancel each other. Between the levels L (free surface of liquid) and M, the force from left is $P_0 hl$, where P_0 stands for the atmospheric pressure, and l for the width of the plate (in a direction perpendicular to the plane of the figure), h being the height of ascent between the plates. The pressure within the raised liquid column at height x above the liquid surface is $P_0 - x\rho g$, where ρ denotes the density of the liquid (g = acceleration due to gravity). This gives the force acting from the right as $(P_0 - \frac{1}{2}h\rho g)hl$.

Thus, the net force acting on the plate A toward the plate B is $\frac{1}{2}h^2\rho gl$ (approx). An estimate for h can be obtained by referring to formula (7-45), which is to be modified by a factor of 2 since the liquid meniscus is curved in only one direction, the curvature in a perpendicular direction

being zero. In other words, $h = \frac{\gamma \cos \theta}{r \rho g}$ (approx), where the notation is as in eq. (7-45). Analogous considerations apply to the case (B), where the pressure at a depth x within the bulk liquid is $P_0 + x\rho g$ while the pressure on the right at the same level is P_0 . This again gives an unbalanced force to the left, of magnitude $\frac{1}{2}h^2\rho gl$, where $h = \frac{\gamma |\cos \theta|}{r \rho g}$, θ being obtuse in this case.

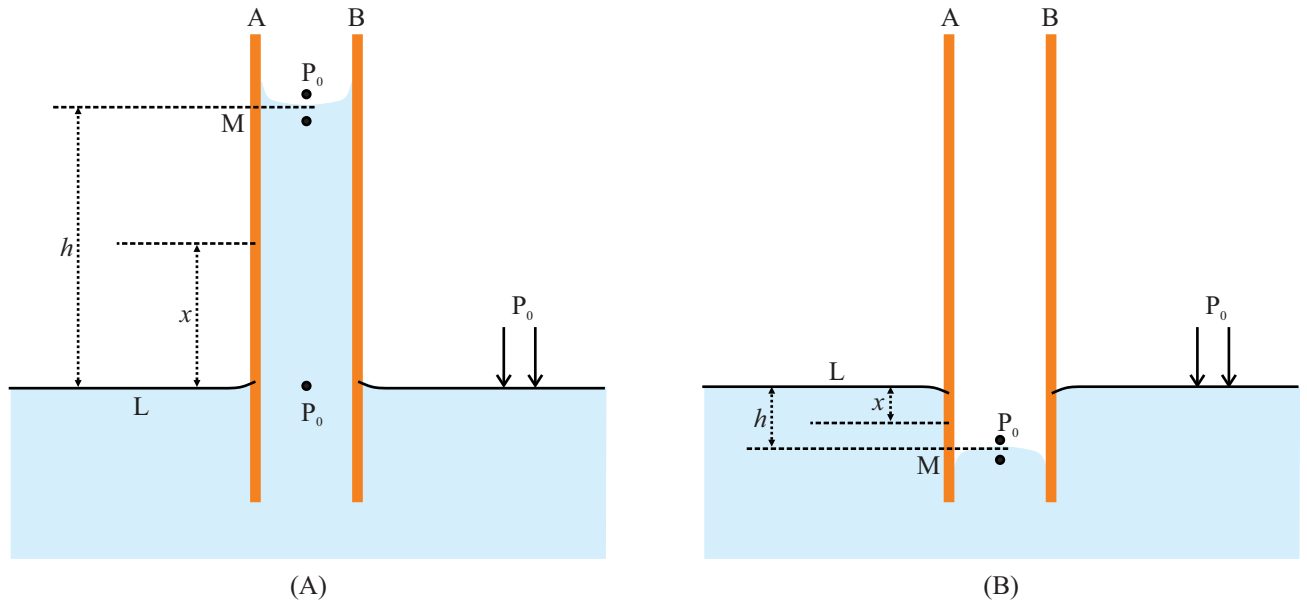


Figure 7-42: Rectangular plates A, B dipped within a liquid, with their planes parallel (schematic); in (A), the liquid rises between the plates up to a height h (level M), owing to an acute angle of contact while in (B), the liquid level gets depressed by depth h , down to M, due to an obtuse angle of contact; L is the free surface of the liquid; P_0 is the atmospheric pressure; the width of either plate in a direction perpendicular to the plane of the figure is l ; a net force acts on the plate A toward B due to the unbalanced forces from two sides in between levels L and M, as in problem 7-12.

Problem 7-13

Water kept in a large trough sticks to the surface of a wall and rises up to a height h as shown in fig. 7-43. Find an expression for the height of ascent h in terms of the surface tension (γ) of water.

Answer to Problem 7-13

HINT: Referring to fig. 7-43, consider the equilibrium of the volume of water within the region ABC (the wall extends in a direction perpendicular to the plane of the figure, being of width, say, l), where A is at the top of the meniscus, and B and C are at the level of the free water surface.

The pressure at any point D at a height x from C is $P_0 - x\rho g$ (ρ is the density of water, g is the acceleration due to gravity, and P_0 stands for the atmospheric pressure, which is also the pressure at C), and hence the horizontal force exerted on the wall is $F = P_0lh - \rho g l \int_0^h x dx = P_0lh - \frac{1}{2}\rho gh^2l$ (reason this out).

The force of reaction exerted by the wall on the water is equal and opposite, being directed toward the right in the figure. The atmospheric pressure P_0 (say) acting on the surface AB exerts a force P_0hl to the left (reason this out; consider the equilibrium of the volume of air bounded by the surface AB and the dotted lines). Finally, imagining a line through B perpendicular to the plane of the figure, running parallel to the width of the wall, a force γl is exerted toward the right by the surface of water to the right of the line.

In other words, one obtains the relation $P_0hl = P_0hl - \frac{1}{2}\rho gh^2l + \gamma l$, giving $h = \sqrt{\frac{2\gamma}{\rho g}}$.

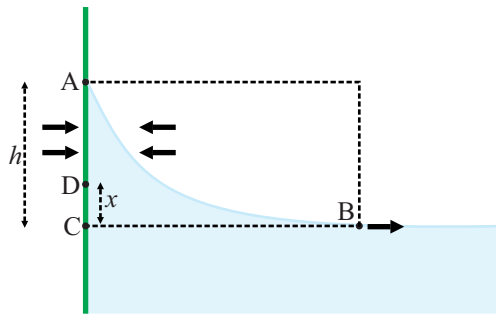


Figure 7-43: Water sticking to the wall of a trough and forming a meniscus of height h (schematic); the point A is at the top of the meniscus, while B and C are at the level of the free water surface; the width of the wall in a direction perpendicular to the plane of the figure is l ; considering the volume of water of width l and bounded within the region ABC, the wall exerts a force on it toward the right while the atmospheric pressure acting on the surface AB exerts a force to the left (this force can be determined by considering the equilibrium of the mass of air of width l , contained in the region bounded by AB and the dotted lines), and the portion of water surface lying to the right of B exerts a pull toward to the right due to surface tension; since the net force has to be zero for equilibrium, the height of the meniscus h is obtained as in problem 7-13.

Problem 7-14

A thin circular layer of water, of diameter D and thickness d , is formed between two parallel glass plates with their planes horizontal. Assuming the meniscus, shown in cross-section in fig. 7-44, to be semicircular, with angle of contact zero, estimate the force between the two plates, assuming

$D \gg d$.

Answer to Problem 7-14

HINT: Referring to the point A within the layer (fig. 7-44), the pressure at any point within the layer is seen to be $P_0 - \frac{\gamma}{d}$, where P_0 stands for the atmospheric pressure and γ for the surface tension of water. Here the meniscus has been assumed to be semi-circular, with radius $\frac{d}{2}$ (of the two principal radii of curvature at A, one is $\approx \frac{D}{2}$ and does not appreciably affect the pressure at A). Thus the upward force exerted by water on the upper plate is $(P_0 - \frac{2\gamma}{d})\frac{\pi D^2}{4}$, while the downward force due to atmospheric pressure is $\frac{\pi P_0 D^2}{4}$. The net downward force, which is also the upward force on the lower plate, is $\frac{\pi \gamma D^2}{2d}$.

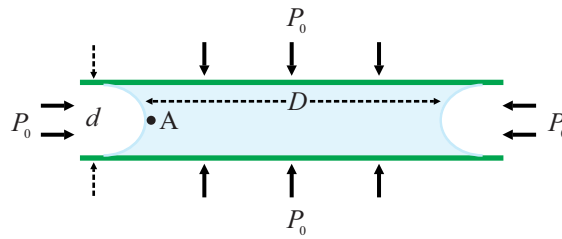


Figure 7-44: Thin circular layer of water between two horizontal flat plates (a cross-section through the center is shown), with a meniscus of semicircular shape and zero angle of contact; the pressure at point A within the layer is less than the atmospheric pressure P_0 by $\frac{2\gamma}{d}$, and can be assumed to be constant throughout the layer; this causes a net force to act on either plate toward the other, as worked out in problem 7-14.

Problem 7-15

Heat of evaporation of a liquid drop.

If L be the latent heat of evaporation (refer to section 8.22.2) from a flat liquid surface, calculate the effective latent heat in the case of evaporation for a drop of radius R of the same liquid, in terms its surface tension (γ) and density (ρ).

Answer to Problem 7-15

HINT: Imagine that, in a certain time interval, a small amount of liquid evaporates from the drop, as a result of which its radius changes from R to $R - \delta R$, the mass of liquid evaporated being

$\delta m = 4\pi\rho R^2\delta R$. In the case of evaporation from a flat surface, the energy δU required for this would come entirely from the internal energy of the liquid, amounting to $L\delta m$. In the case of a spherical drop of radius R , part of the required energy δU comes from the decrease in surface energy, which amounts to $8\pi R\delta R\gamma$ (the decrease in surface area is $8\pi R\delta R$), while the remaining part comes from within the volume of the drop, and is equal to $L'\delta m$, where L' is the effective latent heat for the drop of the given size. In other words, $L\delta m = L'\delta m + 8\pi\gamma R\delta R$, i.e. $L' = L - \frac{2\gamma}{\rho R}$. This turns out to be positive for any liquid drop of even the smallest size (i.e., for one of atomic dimension).

7.6.8 A few phenomena associated with surface tension

Interactions involving interfaces are responsible for a wide variety of natural phenomena and phenomena of daily occurrence. Surface phenomena of various kinds are made use of for industrial and household purposes. We will now have a brief look at a few of these phenomena.

7.6.8.1 Seepage of water through soil

A liquid can easily seep into pores and cavities in a porous medium provided it wets the surface of the material constituting the medium. This property of a liquid is responsible for water seeping through soil and through capillary structures in the roots of plants. The ascent of sap through tubular structures in plants also occurs largely due to the high degree of cohesion among water molecules which is responsible for the high value of the surface tension of water and which prevents the breaking up of the column of water in such a tube.

7.6.8.2 Formation of raindrops and clouds.

The surface tension of a liquid tends to prevent the formation of liquid drops from a bulk assembly of the liquid molecules since it is associated with an energy cost due to the increase in surface area resulting from drop formation.

Problem 7-16

Suppose that a mass m of a liquid of density ρ is to be converted into N number of drops of radius r . Show that the total surface area of the drops will then be

$$S = (4\pi)^{\frac{1}{3}} \left(\frac{3m}{\rho} \right)^{\frac{2}{3}} N^{\frac{1}{3}}. \quad (7-46)$$

Answer to Problem 7-16

HINT: Since the total mass of the N number of drops is to be m , the radius r of each drop is to satisfy the relation $m = \frac{4}{3}\pi r^3 \rho N$; the total surface area $S = 4\pi r^2 N$ is given by (7-46).

Formula (7-46) shows that the larger the number of drops, i.e., the smaller the size of each drop, the larger is the total surface area of the drops formed. Since the surface free energy is proportional to the surface area, the energy cost of formation of drops of small size will be high compared to that for the formation of larger drops.

Surface tension is also relevant in the process of condensation of atmospheric water *vapor* into drops.

In order that water vapor may condense into droplets of water, it is necessary that the atmosphere be *saturated* with water vapor (see sec. 8.22 for basic ideas relating to change of state of materials). However, condensation does not start if the air is just saturated and there are no *nucleation centers* like dust particles, or previously formed drops, present in the air. In the absence of such nucleation centers, the condensation has to begin with drops of infinitesimally small size, for which the energy cost of drop formation will have to be high. In order to make up for the energy cost, the air has to be actually *supersaturated* with water vapor. Such supersaturation is necessary even in the presence of nucleation centers, but then the condensation can start with a relatively low degree of supersaturation.

In summary, the formation of drops of extremely small size requires a high degree of supersaturation of the atmosphere while a lower degree of supersaturation is sufficient for condensation to start in the presence of nucleation centers because then the initial size

of the drops can be larger. Raindrops and clouds are formed around such nucleation centers as the atmosphere attains an appropriate supersaturation level.

The degree of supersaturation necessary for the formation of a drop of radius r at temperature T is expressed in terms of the *saturation vapor pressure* (SVP; refer to section 8.22.4) of a curved surface of the same radius as compared to the SVP over a flat surface. Denoting the two by $P(r)$ and P_0 respectively, the relation between the two is of the following implicit form

$$P(r) = P_0 \exp \left[\frac{M}{\rho RT} \left(\frac{2\gamma}{r} + (P(r) - P_0) \right) \right], \quad (7-47a)$$

where M stands for the molar mass of the liquid, ρ for its density (i.e., $\frac{M}{\rho}$ is the molar volume), and R for the universal gas constant.

In numerous situations of practical interest, the inequality $P(r) - P_0 \ll \frac{2\gamma}{r}$ is satisfied, giving the approximate formula

$$P(r) \approx P_0 \exp \left(\frac{2M\gamma}{\rho RT r} \right). \quad (7-47b)$$

This formula, referred to as the *Kelvin equation*, shows that the SVP for the formation of a liquid drop of radius r increases as the size of the drop decreases.

The derivation of the formula (7-47a) involves thermodynamic considerations relating to phase equilibrium ([9]). The expression for the SVP over a nucleation center or a small liquid drop is of great relevance in the study of colloids and in atmospheric science.

Problem 7-17

Find the relative change in the SVP of water (molar mass $18.0 \times 10^{-3} \text{kg}\cdot\text{mol}^{-1}$, density $1.0 \times 10^3 \text{kg}\cdot\text{m}^{-3}$, surface tension $70 \times 10^{-3} \text{N}\cdot\text{m}^{-1}$) for a drop of radius $1.0 \times 10^{-4} \text{m}$, over the SVP for a flat surface, at temperature $T = 300 \text{K}$ (universal gas constant $R = 8.3 \text{J}\cdot\text{mol}^{-1}\text{K}^{-1}$).

Answer to Problem 7-17

ANSWER: Substituting values in formula (7-47b), and denoting the SVP over the curved surface of the drop by P , one obtains $\frac{P}{P_0} = \exp\left(\frac{2 \times 18.0 \times 10^{-3} \times 70 \times 10^{-3}}{10^3 \times 8.3 \times 300 \times 10^{-4}}\right) \approx 1 + 1.01 \times 10^{-5}$. Thus, the relative change in SVP is $\frac{P-P_0}{P_0} \approx 1.01 \times 10^{-5}$.

NOTE: In this example, $P - P_0 \approx 0.036\text{Pa}$, while $\frac{2\gamma}{r} \approx 1.4 \times 10^3\text{Pa}$, which is why the use of formula (7-47b) instead of the more accurate (7-47a) is justified.

7.6.8.3 Pouring oil over rough sea

Giant waves are formed in the high seas in rough weather due to the impact of strong wind on the waves of relatively smaller size, and also due to the alteration of the tides. On attaining a considerable height, the waves *break up* into great swirling and turbulent masses of water, which are dangerous for navigation. An age-old practice in such situations is to pour oil over the sea water, which has an effect of calming the sea down (this, however, has adverse environmental implications).

It appears that the effect of the oil is to smoothen the surface of the waves by preventing the formation of ripples on these waves and by damping out the ripples already formed. The structure of oil molecules is such that they tend to spread over the surface of water rather than to collect at one place when poured on the water surface. This is why the oil prevents the formation of ripples and dampens them out since the ripples have the effect of disturbing the uniform spreading of the oil film and to cause an increase of surface energy.

As the surfaces of the large waves are smoothened out, the impact of air on the waves becomes less harmful in causing the waves to break. The impact of the wind on the rippled surfaces of the waves is of a cumulative nature, leading to excessive deformation of these waves and their subsequent break-up. The oil film, in a manner of speaking, lessens the 'bite' of the wind on the waves.

A detailed explanation of the action of the oil film in calming the sea involves complex considerations. In this context, you can look up the paper by Peter Behroozi, Kimberly

Cordray, William Griffin, and Feredoon Behroozi [10] entitled ‘The calming effect of oil on water’ (Am. J. Phys., vol 75, p 407-414 (2007)).

7.6.8.4 Walking on water

A number of insects can walk with impunity on water without their body getting wet. The legs of these insects are coated with a waxy secretion so that the water has an obtuse angle of contact with the surface dipping in it (fig. 7-45). The water surface just under the dipping legs then acts like membranous pouches supporting the weight of the insect.

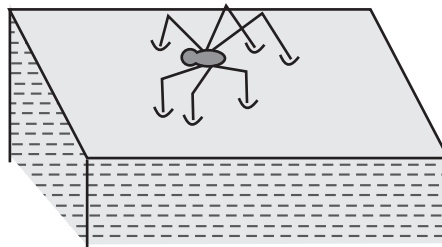


Figure 7-45: Insect walking on water; the legs are covered with a waxy material, and the areas of water in contact with the legs act as membranous pouches, supporting the weight of the insect.

A needle placed carefully on the surface of water floats on it due to a similar reason. Though water wets the surface of the needle, it is possible to place the needle without breaching the water surface, whereupon the needle continues to float since the depressed water surface supports it as a pouch. One can work out the surface tension forces brought to play on the needle and the maximum weight of the needle that these forces can support. Similar calculations can be made for the walking insect as well. The fact that the angle of contact is obtuse in the latter case is an added advantage which explains why the insect does not get wet even though its legs are dipped nearly vertically in the water. By contrast, if the needle is dipped vertically, it sinks because the smaller length of the line of contact between the water surface and the needle results in a smaller surface tension force compared to what is needed to support its weight.

However, recent findings indicate that an alternative mechanism may be involved in

the phenomenon of an insect walking over water. The legs of such an insect are covered with microscopic strands of hairlike attachments indented with extremely small grooves, with air trapped in the grooves and in the space between the strands of hair. These microscopic air bubbles effectively cause the insect to *float* on water by the action of buoyancy and to walk across it.

7.6.8.5 Rayleigh-Plateau Instability: beads on cobweb threads

Imagine a cylindrical jet of liquid, initially of constant radius, falling vertically. As the length of the cylindrical column, in comparison with its radius, increases beyond a certain value, the column breaks up into a wavy structure and finally assumes a periodic form with successive blobs separated by constrictions where the blobs may eventually get detached from the column in the form of a succession of drops.

This phenomenon, first observed systematically by Plateau and subsequently explained by Rayleigh, is based on *instabilities driven by the surface tension of the liquid* developing on the surface of the column, as a result of which small disturbances ('perturbations') arising from various incidental causes get magnified due to mechanisms inherent in the flowing liquid. Any small disturbance on the cylindrical form of the column can be imagined to be made up of a superposition of periodic perturbations of various different amplitudes and wavelengths. Among these, certain perturbations, depending on their wavelengths, and also on the radius of the column, get *amplified* while the rest disappear by getting damped.

Fig. 7-46(A) depicts a periodic perturbation over the cylindrical shape (shown with dotted lines) of the column, where the perturbation is seen to be made up of a succession of blobs (A, C,...) and constrictions (B, D,). Fig. 7-46(B) shows a part of the column in magnified view where it is seen that the geometry of the surface at any point P can be described in terms of two principal radii of curvature - one relating to the circular cross-section of the column in a plane perpendicular to the plane of the figure (shown in perspective by the dotted ellipse), and the other to the cross-section in the plane of the figure (dotted circular arc). The pressure at a point close to the surface just within

the column (point P in the figure) depends on the surface tension of the liquid and on these two radii of curvature, as in formula (7-43).

At a bulge, the radius of curvature in a plane perpendicular to the length of the column gets increased over the radius of the column, thereby tending to diminish the pressure while, at a constriction (point Q), this radius of curvature gets diminished, thereby tending to increase the pressure. As a result, liquid tends to flow at an increased rate *from the constriction to the next lower bulge*, due to which the perturbation gets *amplified*. The effect of the other principal radius of curvature, on the other hand, can be seen to act in the *opposite* direction, tending to cause an increased flow from a bulge to the next lower constriction, thereby dampening the perturbation.

In other words, there appear two opposing influences on the perturbation under question - one tending to amplify it and the other to dampen it. What actually happens, depends on the competition between the two, and is determined by the *wavelength* (λ) of the perturbation under consideration in relation to the radius (R). Rayleigh showed that perturbations with λ larger than $9.1R$ (approx) will grow, while those with a smaller wavelength will get damped. This is corroborated by experimental observations, though the actual dynamics of the column depends in a more complex manner on a number of additional factors.

The basic mechanism of the Rayleigh-Plateau instability is responsible for a host of phenomena, including cosmological ones. As an instance, it is instructive to refer to the curious phenomenon of the formation of arrays of small beads of water droplets along fine cobweb threads due to nocturnal condensation of water vapor. The condensation occurs along the length of a thread, which acts as a nucleation center, but the thin film so formed on the thread is transformed to a periodic array of small beads of water in virtue of the Rayleigh-Plateau instability, though a number of other factors, such as adhesion to the thread, are also of relevance.

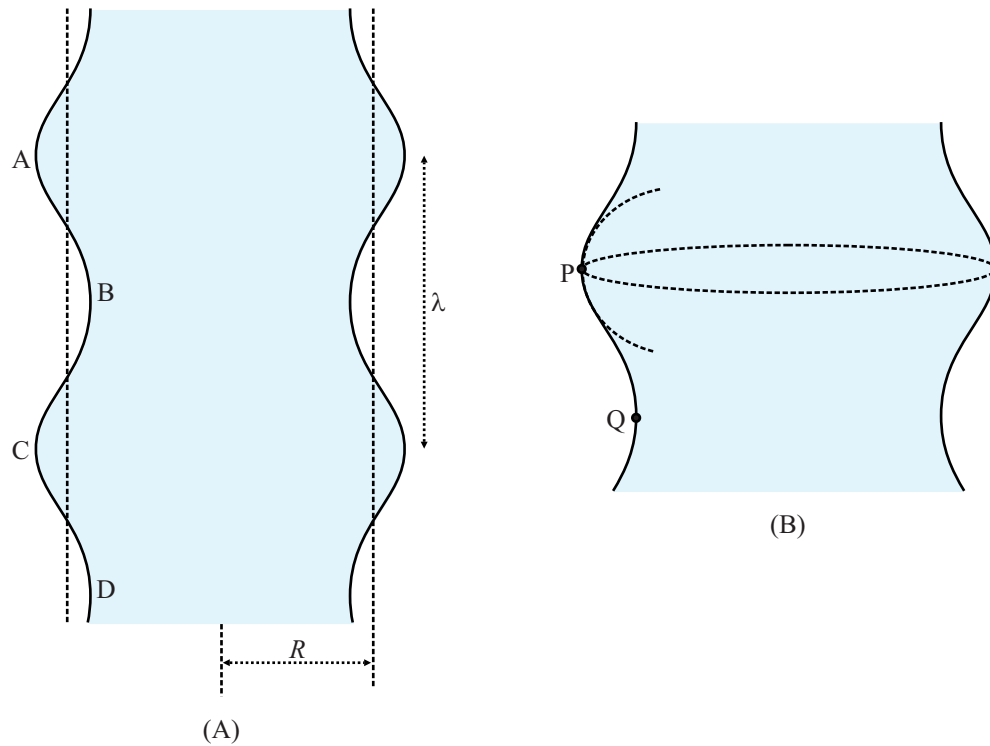


Figure 7-46: Rayleigh-Plateau instability (schematic); (A) a cylindrical column of liquid with a periodic perturbation of the surface; the unperturbed surface is shown with dotted lines; the perturbation is made up of an alternating array of bulges (A, C,...) and constrictions (B, D,...); the dynamics of evolution of the perturbation is determined by the wavelength (λ) in relation to the radius (R) of the column; (B) a bulge and the adjacent constriction, where at any point P close to the surface, the geometry of the latter is seen to be determined by two principal radii of curvature - one in a plane perpendicular to the plane of the figure (depicted by dotted ellipse) and the other in the said plane (circular arc); these two affect the rate of flow from a bulge to the next constriction (or from a constriction to the next bulge) in contrary ways; as a result, some perturbations get damped while others, with $\lambda > 9.1R$ (approx), get amplified; eventually, the column breaks up into a periodic succession of drops.

7.6.8.6 Velocity of surface waves.

The free surface of a large body of water (or any other liquid) in equilibrium under the force of gravity is a horizontal plane. If, due to some disturbance, the surface is made to deviate from its equilibrium configuration, then the subsequent dynamics is determined by an interplay of gravitational and surface tension forces, along with inertial effects, whereby *surface waves* are formed.

Generally speaking, a surface wave (or, more generally, a wave of any other variety, such as an acoustic wave or an electromagnetic wave) can be represented as a superposition

of a group of sinusoidal waves where a sinusoidal wave is characterized by a wavelength λ and an angular frequency ω (related to frequency ν as $\omega = 2\pi\nu$; at times the angular frequency is referred to as the 'frequency' for the sake of brevity) of its own. A sinusoidal member belonging to the group propagates with a *phase velocity* v_p given by

$$v_p = \frac{\omega}{k} \left(k \equiv \frac{2\pi}{\lambda} \right), \quad (7-48a)$$

while the group of waves as a whole has some identifiable structure of its own that propagates with a *group velocity* v_g (see sections 9.15.4 and 14.9 for further considerations relating to group velocity) given by

$$v_g = \frac{d\omega}{dk}. \quad (7-48b)$$

The dynamics responsible for the production of the wave (such as the one involving the gravitational and the surface tension forces in the case of surface waves under consideration) determines the relation between ω and k , known as the *dispersion relation*. In other words, it is the dispersion relation that determines the phase velocity and the group velocity defined above.

In the case of surface waves, gravity is in the nature of a *bulk force* while surface tension results in a *surface force*, where both of these operate to restore the surface of the liquid to the equilibrium configuration, in opposition to the inertial effect that tends to maintain the liquid in motion. For waves with a long wavelength, the gravitational force dominates over the effect of surface tension in determining the dispersion relation, in which case these are referred to as *gravity waves*. The dispersion relation for short wavelengths, on the other hand, is determined principally by surface tension effects, when one refers to the waves as *ripples*.

The general form of the dispersion relation, including both gravity and surface tension effects, turns out to be

$$\omega^2 = gk + \frac{\gamma k^3}{\rho}, \quad (7-49)$$

where g stands for the acceleration due to gravity, γ for the surface tension of the liquid, and ρ for its density.

In the case of a long wavelength gravity wave ($\lambda \gg \sqrt{\frac{\gamma}{\rho g}}$) one obtains

$$v_p = \sqrt{\frac{g\lambda}{2\pi}}, \quad v_g = \frac{1}{2}v_p, \quad (7-50a)$$

while, for a short wavelength ripple ($\lambda \ll \sqrt{\frac{\gamma}{\rho g}}$), one obtains

$$v_p = \sqrt{\frac{2\pi\gamma}{\rho\lambda}}, \quad v_g = \frac{3}{2}v_p. \quad (7-50b)$$

In other words, the phase- and group velocities of ripples increase as their wavelengths are made to decrease, in contrast to the case of gravity waves.

7.6.8.7 Surfactants

Surface active agents (in brief, *surfactants*) have acquired immense economic importance in recent decades. These are chemical substances with a characteristic molecular structure that makes them capable of being adsorbed at the surfaces of various materials and thereby alter profoundly their surface characteristics. These can also form *colloidal* particles that confer useful properties to the liquids in which the colloids are formed.

The way a number of physical characteristics of materials are altered by the addition of surfactants is intimately linked with the chemical structure of the surfactant molecules, and by appropriately manipulating the chemical structures, great versatility can be achieved in the functioning of the surfactants.

A common feature of the large number of surfactant species that have been developed by now is that a surfactant molecule is made of essentially two parts or chemical units - a *lyophilic* part (commonly referred to as the 'head group') and a *lyophobic* part (the 'tail group'), as shown schematically in fig. 7-47. The former of the two is a segment of the surfactant molecule that has an affinity for a certain specific medium while the

latter is a segment that tends to move away from that medium. Since a great majority of surfactants work in a watery medium, the terms *hydrophilic* and *hydrophobic* are commonly used.

When a surfactant is added to water, its molecules get oriented with their head groups surrounded by water molecules at the water surface, and their tail groups sticking away from water (into air or some other medium). The bonding of the water molecules at the surface with the head groups of the surfactant molecules lowers the energy of the water molecules, causing a reduction in the surface tension of water.

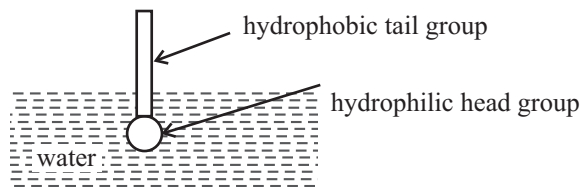


Figure 7-47: Schematic representation of the hydrophilic head group and the hydrophobic tail group of a surfactant molecule; the head group enters into the water while the tail group sticks out.

However, other dispositions of the surfactant molecules in the watery medium are also possible. At low surfactant concentrations, most of the surfactant molecules spread themselves on the surface of water in the manner described above, while relatively few of these molecules remain in the bulk of the watery medium in the form of colloidal particles with a characteristic globular structure called *micelles*. As the concentration of the surfactant molecules is made to cross a certain critical value, corresponding to a *critical micelle concentration* (CMC), there occurs a drastic change in a number of characteristic properties of the system like its surface tension and light scattering efficiency. This transition from the adsorbed surface phase to the micelle phase forms the basis of a number of applications of surfactants.

A notable application of surfactants is in their use as *detergents*. For instance, suppose a drop of oil is to be removed from the surface of a fabric. When an appropriate surfactant solution is applied to the fabric, it seeps through the pores in between the fibers of

the fabric owing to its reduced surface tension and the surfactant molecules get lodged on the surface of the fabric, forming a thin film, *displacing* the oil molecules.

Another important application is based on the ability of a surfactant in *solubilizing* a material in a solvent. For instance, oil, which is immiscible with water, can be solubilized in it by the addition of an appropriate surfactant. Above the critical micelle concentration of the latter, the surfactant molecules aggregate together to form a large number of micelles, with the oil molecules lodged inside the micelles (see fig. 7-48). The micelles get dispersed in water, thereby drawing the oil into solution, where the resulting phase is commonly referred to as a *micro-emulsion*.

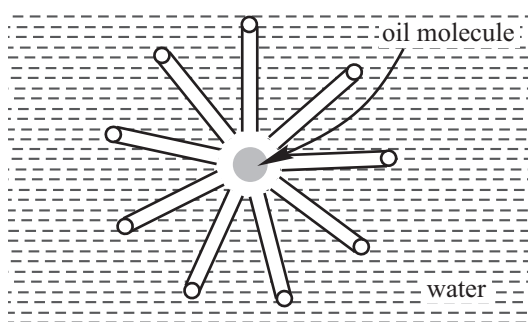


Figure 7-48: Schematic representation of micelle in a solvent with a molecule of an insoluble (or immiscible) substance (say, oil) lodged in its interior; the surfactant molecules are oriented with their head groups towards the bulk of watery medium and their tail groups sticking away from water, while the oil molecule is lodged within the interior of the micelle.

Recent decades have witnessed an unprecedented expansion in the application of surfactants in industry and for household purposes. Surfactants are used in one way or another in textile industry, paper manufacture, leather processing, ore floatation, and electroplating, to name a few of their industrial applications. The use of surfactants is involved in consumer products such as paints, pharmaceuticals, adhesives, cleaning fluids, and cosmetics, including soaps, shampoos, and creams. Indeed, the all-pervasive use of surfactants has now assumed the proportions of a major environmental threat, since many of the surfactants happen to be *non-biodegradable*.

Chapter 8

Thermal Physics

8.1 Thermodynamic systems and their interactions

In thermal physics, we will look at various *thermodynamic systems* and the exchange of energy and matter between them. For instance, one can think of a certain amount of gas in a cylinder fitted with a piston. Or, say, a certain amount of liquid kept in a vessel. These are instances of comparatively simple thermodynamic systems, while more complex systems may also be considered, depending on the context.

I will not attempt to give here a formal definition of a thermodynamic system, and will instead mention that familiar objects of *macroscopic* dimensions around us can be considered as examples of thermodynamic systems. Such objects are made up of a large number of *microscopic* constituents like molecules and atoms. Moreover, they are of a macroscopic *size* compared to microscopic (atomic or molecular) dimensions.

A thermodynamic system can exchange energy and matter with systems external to it. Thermal physics (or, more precisely, *thermodynamics*) deals with the general principles governing such exchanges and allows us to make use of these principles in specific situations.

One can specify or control the manner in which a thermodynamic system exchanges en-

ergy and matter with other systems by keeping it within specific types of *enclosures* or walls. Various different types of enclosures impose different types of *constraints* on the processes of exchange of energy and matter. Among these, I will introduce two particular types of enclosures here and refer to these as *adiabatic* and *diathermic* enclosures respectively.

The terms 'adiabatic' and 'diathermic' are sometimes used in more than one different senses, depending on the context. Possible confusion in this matter will be removed by and by, as we proceed

8.1.1 Adiabatic enclosures: Work

A system enclosed within an adiabatic wall can exchange energy with other systems only in the form of *work*. We assume here that the operational definition of work, including the procedures and principles for measuring work quantitatively are already known to us. For instance, if a rod fitted with vanes be made to rotate in a liquid, some work is done on the liquid, and the amount of work performed can be obtained if the torque applied to the rod and the angle through which the rod rotates are known.

Exchange of *matter* is also possible through an adiabatic wall. Indeed, exchange of matter involves exchange of energy as well. For the time being we will assume that the adiabatic wall allows all types of molecules to pass through it. However, from the point of view of exchange of different types of molecules, an enclosure or a wall can be of various different types. For instance, a wall may be penetrable by molecules of substance A, but may at the same time be impervious to molecules of substance B. Another wall, on the other hand, may be impervious to A while allowing B molecules to pass through. Such walls that allow only certain substances to pass while blocking others are termed *semi-permeable*.

Imagine a chamber divided into two portions by a wall through which the molecules of a particular gas can pass through, but no other form of energy exchange is possible.

Suppose that the gas contained in one portion of the vessel is being made to enter into the other portion by means of a piston which slowly pushes on it so that it is forced through the wall. This requires the performance of some work on the system, which explains how the exchange of matter may involve an exchange of energy in the form of work.

8.1.2 Diathermic enclosures: Heat

In contrast to an adiabatic enclosure, a diathermic one does not allow work to be performed on the system under consideration by other systems. Nor are molecules of matter allowed to pass through it.

I use the term diathermic here in the sense mentioned above. In some contexts, the term diathermic may be meant to refer to an enclosure that allows specific types of molecules to pass through it.

The energy imparted to the system under consideration from other systems through such a wall will be referred to as *heat*. A few more basic concepts are, however, necessary before we come up with a quantitative definition of heat.

8.1.3 Examples of adiabatic and diathermic enclosures

Can the adiabatic and diathermic walls referred to above be found in real life? The way I have introduced these, these appear as idealized concepts. However, one can identify substances that make up walls or enclosures with properties close to the ideal ones. For instance, in the case of a gas kept in a cylinder made of a rigid metallic material, the wall of the cylinder may be assumed to be a diathermic one. Such a wall prevents energy exchange in the form of work, as also the energy flowing by means of exchange of matter. Work by means of electrical and magnetic influence can also be minimized through appropriate measures. It is the energy that *can* still be exchanged through this wall, that is referred to as heat.

Examples of adiabatic walls are more difficult to cite. Think of a gas kept in a cylinder

whose wall is made of a rubber-like material, the cylinder being fitted with a frictionless piston made of the same material. In this case the exchange of energy with other systems will take place only in the form of work performed on or by the gas. Such work can be performed by moving the piston inward or outward, or even by magnetic or electrical influence exerted from outside. Moreover, as mentioned above, another type of work is to be reckoned with, namely, through the exchange of matter. Now, rubber is not easily permeable to molecules of various substances. If the wall is made of some *porous* material like, say, cotton or wool, then it becomes permeable to molecules of different kinds. A number of artificially synthesized substances may also be used as materials for permeable adiabatic walls.

8.2 Thermodynamic equilibrium

Think of a thermodynamic system enclosed in a wall of some specific type, where it is involved in an interaction (or exchange), also of some specific kind, with external systems.

For instance, one can think, in particular, of an *isolated* system. The enclosure in this case will have to be made of a combination of a diathermic and an adiabatic wall which will prevent *all* energy and matter exchange between the system under consideration and other, external, systems. Strictly speaking, though, no system can be *completely* isolated from the external world since, in practice, some weak and residual influence of the latter will always be found to remain. In spite of this, however, we will refer to the system as an 'isolated' one.

On observing such an isolated thermodynamic system for a sufficiently long time under given conditions, one finds that the system ultimately arrives at some *equilibrium* state. If subsequently one measures at different points of time some physical characteristic or other pertaining to the system as a whole (rather than to one or more of its microscopic constituents) one would find that all the measured values are the same (i.e., time-independent). This means that, in a macroscopic sense, the system has reached a constant state. Looked at from the *microscopic* point of view, however, the character-

istic physical quantities pertaining to the individual constituents may go on changing quite rapidly. For instance, even when the state of the gas in the cylinder considered as a whole becomes constant and unchanging, the states of motion of the individual molecules of the gas may continue to change.

In summary, in a state of equilibrium all the physical characteristics of the system considered as a whole have well defined time-independent values. These are termed the *thermodynamic variables* or *thermodynamic state functions* ('thermodynamic functions' or 'state functions' in brief; the term 'state variables' is also used) for the system. Stated in a slightly different way, a set of well-defined values of the thermodynamic variables identifies uniquely a state of thermodynamic equilibrium.

For instance, in the above example of a gas kept in a cylinder, variables like the volume, pressure, temperature and specific heat constitute examples of thermodynamic variables (also referred to as *state variables*) of the gas. All of these have well defined values in any given equilibrium state of the gas. However, all these are not mutually independent variables. For instance, for a given quantity of the gas, assumed to be a pure one, if one knows the values of pressure and volume of the gas, one can work out the values of all the other state variables. In other words, for a given quantity of a pure gas, only two appropriately chosen state variables specify uniquely an equilibrium state. For a *mixture* of several gases, on the other hand, additional parameters are necessary for a complete specification of an equilibrium state.

8.3 Thermodynamic processes

Starting with a thermodynamic system in some definite equilibrium state (hereafter, we will refer to an equilibrium state as, simply, a *state*, provided, of course, there is no scope for confusion), one can disturb the system by means of some influence brought to bear on it by appropriate means, and bring it to a *new* equilibrium state. The changes taking place in the system in between the initial and final equilibrium states, is then said to constitute a *thermodynamic process*.

8.3.1 Adiabatic process

An instance of a thermodynamic process is an *adiabatic* one. Any change taking place in a system enclosed within an adiabatic enclosure is termed an adiabatic process. The system can exchange energy with the outer world in the form of work in such a process, but not in the form of heat. Additionally, the quantity of one or more components in the system may also change, either by exchange of material with the external world or through some chemical reaction or other taking place within it.

Think of a fixed amount of gas in a cylinder fitted with a piston, kept in an adiabatic enclosure where the enclosure is assumed to be an impermeable one. Suppose that initially the gas is in an equilibrium state with pressure p_1 , with the volume of the gas being V_1 . Let now the piston be moved to a new position where the volume of the gas becomes V_2 while the pressure, after attainment of equilibrium in this new position of the piston, becomes p_2 . The entire process from the initial position of the piston to the final position is then an instance of an adiabatic process.

8.3.2 States and processes: thermodynamic state diagram

If the volume (V) and pressure (p) in any equilibrium state of the gas in the above example be indicated by the x- and y- co-ordinates of a point in a graphical diagram then that point may be taken to represent the state under consideration in the given context. Such a diagram is referred to as a thermodynamic *state diagram* for the system. In other words, a state diagram is a graphical representation of equilibrium states of a system with the help of points in the diagram. A state diagram may involve more than two co-ordinates, depending on the number of independent state variables of the system under consideration.

I should mention that one can represent only *equilibrium* states in a state diagram. A state of the system which is not time-independent, i.e., one for which the values of the state variables keep on changing with time (these are referred to as *non-equilibrium* states) have no place in the state diagram. The question then comes up as to whether

the states in the intermediate stages of a process starting from one equilibrium state and ending up in another, can be represented in the state diagram.

8.3.3 Quasi-static processes

Think of a thermodynamic process that is being made to occur at a very slow pace. Suppose that the pace is so slow that at every intermediate stage of the process, the system attains some equilibrium state or other. For instance if the piston be made to move very slowly in the above example of an adiabatic process in a gas enclosed in a cylinder, then at every intermediate stage the gas attains some particular volume (V) and pressure (p), depending on where the piston has moved to at that stage, and every such intermediate state of the gas can then be represented by a point in the state diagram. In other words, one will then have a continuous succession of intermediate equilibrium states in between the initial and final states in the diagram. These will then constitute a continuous *path* in the state diagram connecting the initial and final states. Such a slow process in a system that can be depicted by a continuous path in the state diagram is referred to as a *quasi-static* process.

Fig. 8-1 depicts the p - V state diagram of a fixed quantity of a pure gas, in which a path corresponding to a quasi-static adiabatic process has been shown schematically.

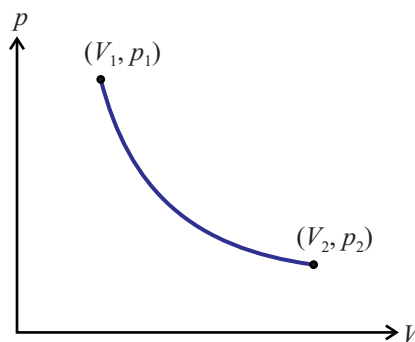


Figure 8-1: State diagram of a fixed quantity of a pure gas; the path in the state space corresponding to a quasi-static adiabatic process has been depicted schematically; the points (V_1, p_1) and (V_2, p_2) correspond to the initial and final states of the gas.

Note that if the process were made to occur at a *rapid* pace then the gas could not

have attained equilibrium in any intermediate stage and then it would not be possible to represent the adiabatic process by a path since only the initial and final states could then be depicted in the diagram.

8.3.4 Thermal equilibrium

Fig. 8-2(A) and (B) show the state space diagrams of two thermodynamic systems A and B that can interact with one another through a diathermic wall, where the composite system made up of A and B is isolated from the rest of the world. One then says that a thermal contact has been established between A and B. With this thermal contact between the two, as the systems A and B attain equilibrium, they are said to be in *thermal equilibrium*. Recall that, by the definition of a diathermic wall, no exchange of energy between A and B can take place in the form of work, where the latter includes work done by the exchange of material components. Suppose that the points a_1 and b_1 in figs. 8-2(A) and (B) respectively represent the states of the two systems in thermal equilibrium. For instance, if A and B be fixed quantities of two gases, then the states can be depicted in a p - V state diagram for each gas, as in the figure. One says that the states a_1 and b_1 of A and B respectively are in thermal equilibrium.

8.4 The zeroth law of thermodynamics: Temperature.

8.4.1 Explaining the zeroth law

Think now of any three thermodynamic systems, say, A, B, and C. Let the states a_1 and b_1 of A and B respectively (shown in fig. 8-2(A), (B)) be related by thermal equilibrium. Again, let the state b_1 of B be in thermal equilibrium with state c_1 of C (not shown in fig. 8-2). Then the *zeroth law of thermodynamics* states that the states a_1 of A and c_1 of C will also be related by thermal equilibrium. In other words, the relation of states of systems being in thermal equilibrium with one another is *transitive*. This is the substance of what has come to be known as the *zeroth law of thermodynamics* - a principle arrived at on the basis of observations on states of systems in thermal equilibrium with one another.

It may be noted that states of the *same* system can also be related to one another by thermal equilibrium. For instance, in fig. 8-2(A), the states a_1 , a_2 , and a_3 of the system A are all in thermal equilibrium with one another (this can be interpreted as all these states being in thermal equilibrium with some given state of another system, or as states of three *copies* of the same system A being in thermal equilibrium with one another).

Again, in figure 8-2(B), the states b_1 , b_2 , and b_3 of B are in thermal equilibrium with one another, as also with the states a_1 , a_2 , and a_3 of A (note that the zeroth law is operative here). In this manner, one can think of states of other systems such as, say C, D, etc., such that sets of states of these systems are also in thermal equilibrium with the states a_1 , a_2 , and a_3 of A, as also with states b_1 , b_2 , b_3 of B.

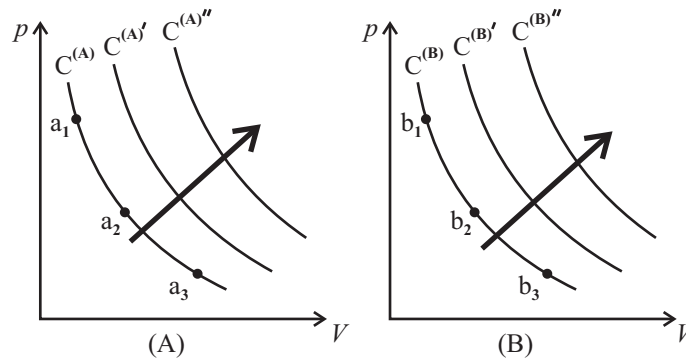


Figure 8-2: Explaining the idea of thermal equilibrium and of the zeroth law of thermodynamics; in (A), the states a_1 , a_2 , a_3 of the system A are in thermal equilibrium with one another while similarly, in (B), the states b_1 , b_2 , b_3 are in thermal equilibrium with one another as also with the states a_1 , a_2 , a_3 of A; a curve $C^{(A)}$ is drawn in (A) with all states of A that are with thermal equilibrium with any of the three states a_1 , a_2 , and a_3 lying on it, the corresponding curve in (B) being $C^{(B)}$; these two then constitute *isothermal curves* for A and B corresponding to the same temperature; other such pairs of curves $C^{(A)'}$, $C^{(B)'}$, and $C^{(A)''}$, $C^{(B)''}$ are also shown; the arrows indicate a monotonic change in the temperature.

8.4.2 Temperature as a thermodynamic variable

Using in this manner the transitive property of the relation of thermal equilibrium of states, we can think of sets of states of all thermodynamic systems such that all are in thermal equilibrium with one another. This suggests that all these states must have

some characteristic in common, i.e., there must be some common state variable whose value is the same for all these states. This state variable we call *temperature*.

While this is not a logically sound derivation of the concept of temperature starting from the zeroth law, more detailed considerations show that the existence of a state variable - the temperature - does indeed follow from the zeroth law, where the variable possesses the same value for all states in thermal equilibrium with one another, regardless of the systems for which these states are defined.

Thus, all states of the system A lying on the curve $C^{(A)}$ in fig. 8-2(A) as also states lying on the curve $C^{(B)}$ in fig. 8-2(B), correspond to the same value of the temperature. These two then constitute what are referred to as *isothermal* curves for the two systems under consideration, corresponding to the same value of the temperature (other systems like C, D, etc., not represented in fig. 8-2(A), (B), may also be considered along with A and B).

8.4.3 Empirical scales of temperature

Considering the isothermal curves $C^{(A)}$ and $C^{(A)'}$ in fig. 8-2(A) (or $C^{(B)}$ and $C^{(B)'}$ in fig. 8-2(B)), it can be seen that no state on $C^{(A)}$ is in thermal equilibrium with any state on $C^{(A)'}$, i.e., the isothermal curves are all disjoint, and so it is possible to label these curves uniquely with numbers such that the numbers change monotonically as the curves are crossed in succession in any given direction, say, the one shown by the arrows in fig. 8-2(A) and (B), it being understood that corresponding curves for other systems (such as, say, C or D) are also labeled with the identical set of numbers.

This, then, will constitute the consistent assignment of a unique value of temperature for each state of any and every thermodynamic system, where the possibility of such a consistent assignment of values is in conformity with the fact that temperature is a state variable of any thermodynamic system.

In reality, there exists not one but an *infinite* number of ways in which such a consistent

assignment of values is possible. The choice of any one of these possible assignments of temperature values gives an *empirical scale of temperature*.

In practice, an empirical scale of temperature is realized by choosing a reference system, called a *thermometer*, and considering an appropriate state variable relating to that system, called its *thermometric property*. For instance, one may choose as the reference system a fixed quantity of a gas maintained at a given pressure. The volume of the gas in various different states can then be taken as the thermometric property. If the volume be V for any given state of the gas, then this can be taken as a measure of the temperature of any system in any state in thermal equilibrium with that particular state of the thermometer. Various different values of V (for the fixed value of the pressure under consideration) will then correspond to different temperatures. Any appropriately chosen *monotonic* function of V may also be taken to constitute an empirical temperature scale.

Analogously, a wire of pure platinum may constitute the thermodynamic system constituting the thermometer and the electrical resistance (R) of the wire in any given state may be taken as the thermometric property. Considering any other system in a state in thermal equilibrium with the platinum wire, the temperature of that system can then be expressed in terms of R , or an appropriately chosen function of it. This then constitutes another empirical temperature scale.

The commonly used centigrade (or Celsius) and Fahrenheit scales are two such empirical temperature scales, where the thermometric property is the length of a column of mercury in a narrow closed tube, with a bulb containing mercury at one end of the tube.

Though one can express the value of temperature in various possible temperature scales that differ from one another, the values of temperature for a given equilibrium state of a system on these different scales must bear a definite relationship between themselves because they represent the same physical quantity. To put it differently, the scales have to be *compatible* with one another.

For instance, if the value of a given temperature in the Celsius scale be C and the value

of the same temperature in the Fahrenheit scale be F , then the two are related as

$$\frac{C}{5} = \frac{F - 32}{9}. \quad (8-1)$$

Another temperature scale of great importance from the scientific point of view is the *thermodynamic* or *absolute scale* (refer to sec. 8.15.4 below; see also sec. 8.12). The definition of the absolute scale of temperature is not tied to any particular thermometric property of any particular thermodynamic system, but is based on a *universal* property of all systems, relating to the *second law of thermodynamics* (see sec. 8.13). However, this definition refers to a set of *idealized* processes occurring in a system and as such, there remains the question of the *practical realization* of this scale of temperature i.e., an operational procedure for the determination of the temperature in the thermodynamic scale of any given system in any chosen state.

The *standard international* (SI in brief) scale, or the *Kelvin* scale of temperature (see sec. 8.4.4 below) constitutes such a practical (though, strictly speaking, an approximate) realization of the absolute thermodynamic scale. If the value of a temperature expressed in the Celsius and the Kelvin scales be C and T respectively, then one has

$$T = C + 273.16. \quad (8-2)$$

For our purpose, in considerations of a theoretical nature, the thermodynamic scale is invoked though, whenever numerical values are to be assigned in a theoretical formula, the Kelvin scale is made use of. The thermodynamic scale, in its turn, is defined in a way so as to make it consistent with another scale of temperature arrived at from idealized considerations, namely, the absolute gas scale (or, in brief, the *gas scale*; see heat1-subsec16). In this book, we will not distinguish between the thermodynamic scale, the Kelvin (or SI) scale, and the gas scale. You will find a summary of considerations relating to the various scales of temperature relevant to thermodynamics in sec. 8.15.5 later in this chapter.

8.4.4 The SI scale of temperature

According to the principles of thermodynamics, there is a unique temperature such that no system can be cooled to a temperature below this. The SI scale is defined by assigning the value 0K to this temperature, referred to as the *absolute zero*. A state of any system at absolute zero is characterized by zero value of the *entropy* of the system (see sections 8.12, 8.14) as also of its specific heats (see sec. 8.21.2), measured in any empirical temperature scale. In addition to the absolute zero, the SI scale of temperature is defined with reference to another *fixed point* (a reference point assigned a definite value in establishing a temperature scale), namely the *triple point of water* (see sec. 8.22.5), for which the temperature is taken to be 273.16 K. The unit of temperature difference, the kelvin degree, is defined to be $\frac{1}{273.16}$ times the difference of temperature between the triple point of water and the absolute zero.

The operational procedure for the measurement of any given temperature in the SI scale involves the use of a specified set of standard thermometers (e.g., the *thermocouple* thermometer and the *platinum resistance* thermometer) in specified temperature ranges.

More recently, it has been proposed to redefine the kelvin degree in terms of the *joule* (the unit of energy) by assigning a certain specified value to the *Boltzmann constant* (k_B ; see sec. 1.4.3). Indeed, from a fundamental point of view, temperature is a physical quantity that can be expressed in energy units. The fact that temperature and energy are expressed in terms of two different scales is a matter of convention.

8.4.5 The direction of heat flow

The zeroth law of thermodynamics makes possible the assignment of a unique value of temperature for every state of any given thermodynamic system such that two systems in states with the same temperature are in thermal equilibrium. Conversely, one can state that systems at different temperatures cannot be in thermal equilibrium. If two systems at different temperatures are made to interact through a diathermic wall, their states undergo a process of change, whereafter a condition of thermal equilibrium is established between them.

What happens in the latter situation is that a flow of *heat* occurs from one of the systems to the other. The concept of heat is defined quantitatively with reference to the *first law of thermodynamics* (see sec. 8.7).

While the zeroth law implies that a flow of heat takes place between two bodies maintained at different temperatures, it does not specify the *direction* of that flow. It is the *second law of thermodynamics* (see sec. 8.11) that specifies the direction of the heat flow. As we know from experience, energy in the form of heat flows from bodies at higher temperatures to ones at lower temperatures when they are made to be in thermal contact. The second law tells us that there is a fundamental principle involving macroscopic bodies underlying this fact of experience, namely the *entropy principle* (briefly introduced in sec. 8.11) which, in essence, is one version of the second law of thermodynamics.

8.4.6 Thermal reservoir: heat source and heat sink

Think of a thermodynamic system S having a large volume and mass. Suppose that the temperature of S in one of its equilibrium states is T and that it is in thermal contact with a system A , the latter having a mass and volume small compared to those of S . As I have already mentioned, the thermal contact is established by letting S and A interact through a diathermic boundary wall. If, now, the initial temperature of A be less than that of S then heat will flow from S to A till the system A attains the temperature T . However, the system S being a very *large* one, its temperature will not be altered much in the process. Indeed, assuming the volume and mass of S to be sufficiently large, one can ignore the change in temperature of S . Consequently, when thermal equilibrium is established between S and A , it will be found that the temperature of S continues to be at T while that of A also reaches the value T . One then says that, with respect to A , S acts as a heat *source*.

If, on the other hand, the initial temperature of A be higher than T , then heat will flow from A to S till the temperature of A reaches the value T , while the temperature of S will continue to remain at the value T . S is then said to be a heat *sink* with respect to A .

A large system that may act as a heat source in some situations and a heat sink in

some others, is termed a *thermal reservoir* (or heat reservoir). In practice, the mass of air surrounding a system often acts as a thermal reservoir for it.

8.4.7 Isothermal processes

If, at every stage of a thermodynamic process involving a given system, the temperature of the system remains constant, then the process is said to be an *isothermal* one (relative to the system under consideration).

Think of a given quantity of a gas kept in a cylinder fitted with a freely moveable piston. Assume that the walls of the cylinder are made of diathermic material, while the piston is in the nature of a moveable adiabatic boundary, but one impermeable to material particles. Thus, in the present instance, there are two distinct ways that energy can enter into or leave the gas contained within the cylinder: as heat flow through the diathermic walls, and as work performed on or by the gas by means of the piston. Suppose further that the heat exchange takes place with a thermal reservoir surrounding the cylinder maintained at temperature, say, T (see fig. 8-3).

In this instance, if the piston is moved quasi-statically i.e., at an infinitesimally slow rate, then gas will reach a state of thermal equilibrium with the reservoir at every intermediate stage of the process, and so the temperature at every stage remains constant at the value T . This then constitutes an example of an *isothermal* process undergone by the gas.

Notice that, during every stage of an isothermal process, the system under consideration has to be necessarily in a state of thermal equilibrium, which means that an isothermal process has to be a *quasi-static* one. As a consequence, it can be represented by a continuous path in a state diagram. By contrast, an adiabatic process need not be quasi-static, and it may not be possible to represent it by a path in a state diagram.

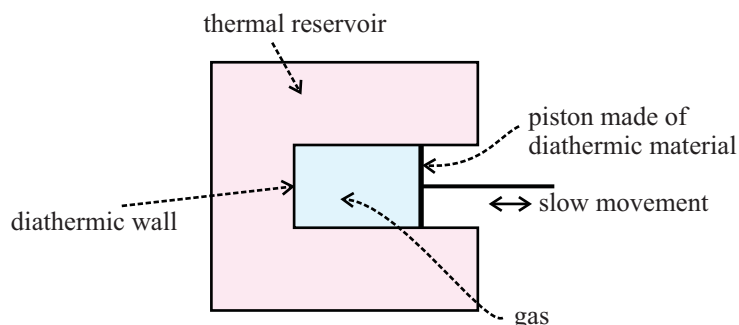


Figure 8-3: Illustrating an isothermal process; while the piston is shown to be made of a diathermic material, it is more appropriately to be one made of an adiabatic material if the system on the other side of it is one other than the thermal reservoir, in which case, no heat exchange can take place other than that with the reservoir.

8.5 Adiabatic processes between given states

Recall that, while the state of a fixed quantity of a gas can be represented by a point in a two dimensional state diagram (with pressure and volume making up the two dimensions), more generally one may require a space of more than two dimensions in which a state of a given system can be represented by a point. For the sake of illustration, however, we will continue to refer to two dimensional state diagrams.

Think of any given thermodynamic system and imagine two points, say, 'a' and 'b', in the state space of the system, representing two of its equilibrium states. The question that now comes up is, can the system be taken from state 'a' to state 'b' by an adiabatic process? Note that the states 'a' and 'b' have been chosen *arbitrarily*, without any pre-determined criteria.

On the face of it, one might think that an adiabatic process may not always be found taking the system from 'a' to 'b'. Suppose, for instance, that there exists an isothermal process taking the system from 'a' to 'b'. Recalling that the system exchanges heat with the external world in an isothermal process, how can it be possible to take the system from 'a' to 'b' by an adiabatic process where *no* exchange of heat can take place?

In reality, however, an adiabatic process can *always* be found that takes the system from 'a' to 'b', regardless of our choice of the two states. For instance, a process where heat is

supplied to a gas at constant volume (causing an increase in pressure and temperature) can be replaced by one where a rod with vanes is made to rotate in the gas, kept at a constant volume. Here the rotating rod with the vanes performs work on the gas, but no heat is supplied to it, though the effect of the process is the same as if heat were supplied from outside.

However, an adiabatic process that has the desired effect of taking the system from 'a' to 'b' *need not be a quasi-static one*, which means that it may not be possible to represent it with the help of a path in the state space of the system. For instance, in the above example of work being done by the rod with the rotating vanes, the flow of the gas around the vanes does not allow the gas to attain equilibrium during the process.

Another relevant fact pertaining to adiabatic processes is that, for any two given equilibrium states 'a' and 'b' of a system, it is usually possible to have not one but *several* (in general, an *infinite* number of) adiabatic processes taking the system from state 'a' to state 'b'. For instance, in the example of the gas and the rotating vane, a different adiabatic process connecting the states 'a' with 'b' would correspond to a different speed of rotation of the vane. Fig. 8-4 depicts two adiabatic processes represented by dotted lines because these processes are not, in general, quasi-static.

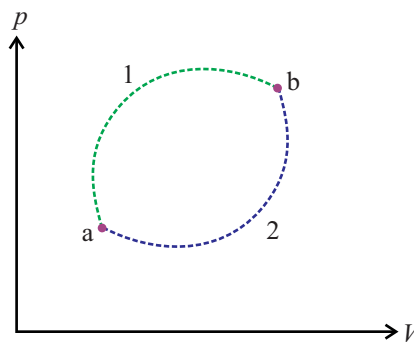


Figure 8-4: Depicting two possible adiabatic processes connecting states 'a' and 'b'; the processes (marked '1' and '2'), are shown with dotted lines in the state space since they need not be quasi-static and hence, and hence do not occur through successions of equilibrium states.

Recall that in an adiabatic process the system under consideration exchanges energy

with other systems in the form of work but not in the form of heat (I will present the quantitative definition of heat in a little while). If, during any stage of the process, external systems perform W amount of work on the system, then one can state equivalently that work performed *by* the system on these external systems is $-W$. In other words, work performed *on the system* and that *by the system* are related to each other by a negative sign.

8.6 The significance of adiabatic work

The work done in an adiabatic process is sometimes referred to as *adiabatic work* in brief.

As we have seen, there may be, in general, *more than one* adiabatic processes taking a given system from any state 'a' to another state 'b'. We assume that no part of the work done on the system in any such process goes to change the translational or rotational kinetic energy of the system as a whole. It has been found experimentally that the following statement holds for all such processes:

If a system is taken from a state 'a' to another state 'b' by more than one adiabatic processes, then the work performed on that system by external systems will be the same for all such processes.

This, indeed, is the content *first law of thermodynamics*, though the first law is commonly stated in a different form. I will later present a modified form of the above statement that is closer to the familiar form of the first law.

The above statement implies that, for any adiabatic process taking the system under consideration from state 'a' to state 'b', the work done on the system (or, equivalently, the work done *by* it), does not depend on the process chosen but depends solely on the two states 'a' and 'b'.

Let us denote the amount of this work done on the system in an adiabatic process taking it from 'a' to 'b' as W_{ab} . Think now of any chosen reference state of the system and call it

'r'. One can chose the reference state according to convenience, but once the state has been chosen, no *other* state can simultaneously be used as the reference state. With this reference state 'r', imagine a process that takes the system from 'r' to 'a', and then from 'a' to 'b'. According to the notation introduced above, the amount of adiabatic work in the first part of the process is W_{ra} while that in the second part is W_{ab} . Hence the total amount of adiabatic work done on the system in the entire process initiated at 'r' and terminated at 'b' is $W_{ra} + W_{ab}$. Evidently, then, one can write

$$W_{rb} = W_{ra} + W_{ab}, \quad (8-3a)$$

which implies

$$W_{ab} = W_{rb} - W_{ra}. \quad (8-3b)$$

If one now thinks of states such as, say, 'a', 'b', 'c', ..., and the associated quantities W_{ra} , W_{rb} , W_{rc} , ..., then the reference state 'r' being common to all these quantities, one may think of these as depending solely on the states 'a', 'b', 'c',... respectively and denote these as, say, $U(a)$, $U(b)$, $U(c)$,.... . These are then referred to as the *internal energy* of the respective states 'a', 'b', 'c', One thereby arrives at the concept of *internal energy* (commonly denoted by the symbol U) as a definite thermodynamic variable because, as we have just seen, once the state 'r' is fixed, U depends only on the state of the system under consideration.

It is not difficult to have an idea of the physical significance of this thermodynamic variable U . Notice from (8-3b) that the work done on the system in the adiabatic process 'a'→'b' is given by

$$W_{ab} = U(b) - U(a). \quad (8-4)$$

Since, in this process, energy is exchanged between the system and the outer world only in the form of work, one can say that the amount of energy flowing into the system in the process under consideration is $U(b) - U(a)$. If one now interprets $U(a)$ as the sum of

potential and kinetic energies of all the constituents of the system taken together in the state 'a', and similarly $U(b)$ as the total kinetic and potential energy of the constituents in the state 'b', then (8-4) is seen to be amenable to a consistent interpretation, because it then tells us that the energy flowing into the system in the process 'a'→'b' is the difference of total energies of the constituents of the system in the states 'b' and 'a', as it should be. Indeed, this is consistent with the *principle of conservation of energy*.

In other words, the internal energy as defined above, is simply the total energy of the constituents of the system under consideration in any given state. The unit of internal energy in the SI system is thus seen to be *joule* (J).

Note that if, instead of the reference state 'r', one chooses any other state, say, 'r'' as the reference state then the internal energy for the state 'a' will be given by

$$U'(a) = W_{r'a} = W_{r'r} + W_{ra} = U(a) + W_{r'r}, \quad (8-5)$$

instead of $U(a)$. The difference of internal energies of any given state, defined with 'r'' and 'r' as reference states, is thus seen to be a constant independent of the state under consideration, i.e., is the same for all states 'a', 'b', 'c', In other words, the internal energy is not a uniquely defined quantity, but is arbitrary only to the extent of an additive constant. This does not indicate any inconsistency in our interpretation and is simply a reflection of the fact that there is an arbitrariness of an additive constant in the definition of energy itself. For instance, we saw in chapter 3 that the potential energy of a system of particles has a similar arbitrariness in it, while the kinetic energy can also be redefined with a constant added to it.

In the above considerations, it has been implicitly assumed that the work done *on* a system in a process is of the same magnitude as the work done *by* it, while being of the opposite sign. Considering two systems A and B, the work (W_{AB}) performed by B on A will be related to the work (W_{BA}) performed by A on B as $W_{AB} = -W_{BA}$ provided the interaction between the two systems is set up in an appropriate manner.

Problem 8-1

The work done on a system in an adiabatic process taking it from a state A to state B is 20.0J, while the work done by the system in another adiabatic process taking it from state B to state C is 10.5J. If the internal energy of the system in the state C is 30.2J, find the internal energy in the state A.

Answer to Problem 8-1

SOLUTION: Following the notation in formula (8-4), we have $U_B - U_A = 20.0\text{J}$. The work done on the system in the process from B to C being -10.5J , we have $U_C - U_B = -10.5\text{J}$, which gives $U_C - U_A = 9.5\text{J}$; thus, $U_A = (30.2 - 9.5)\text{J}$, i.e., 20.7J.

8.7 The quantitative definition of heat

The above considerations lead us to conclude that heat and work are two distinct modes of energy exchange between a system and its surroundings. When the system is kept in a diathermic enclosure, energy exchange in the form of work is prevented and the only mode of exchange then is heat. This is in contrast to an adiabatic process in which the allowed mode of energy exchange is work.

In reality, there may occur very many types of changes in a system *other* than these two special types, in which *both* heat and work are exchanged with the surroundings, the isothermal process considered in sec. 8.4.7 being an instance. Imagine a change of such a more general type to take place in the system under consideration, in which it passes from an initial state 'a' to the final state 'b'. Suppose that the work done on the system in this process is W . Now imagine a *second* process, an *adiabatic* one, that takes the system from the same initial state 'a' to the same final state 'b', it being, as we know by now, always possible to design such a process for any given pair of states. Supposing that the work done on the system in this adiabatic process is W_{ab} , one has equation (8-4) connecting W_{ab} with the internal energies $U(a)$ and $U(b)$ in the states 'a' and 'b' respectively.

Now, since the internal energy is a thermodynamic variable, *its change in the first process must also be* $U(b) - U(a)$, i.e., W_{ab} , which, according to our interpretation, has to be the total energy flowing into the system in the first process. This need not be the same as the amount of energy (W) flowing into the system in the form of work in the first process since some energy may flow into it in the form of *heat*. For the sake of consistency, then, the quantity of heat flowing into the system in the first process which, in the present context, is an arbitrarily chosen one, has to be defined as

$$Q = U(b) - U(a) - W (= W_{ab} - W). \quad (8-6)$$

8.8 The first law of thermodynamics

Thus, in any given process, not necessarily isothermal or adiabatic, the quantity of heat entering into the system is given by

$$Q = \Delta U - W, \quad (8-7a)$$

where ΔU denotes the change in internal energy of the system in the process (measured by the work done on the system in an imagined adiabatic process connecting the relevant initial and final states). In the context of the process under consideration, involving an exchange of both heat and work, ΔU thus appears in the form

$$\Delta U = Q + W, \quad (8-7b)$$

which constitutes the *mathematical formulation of the first law of thermodynamics*. Notice that it is nothing but the principle of conservation of energy, written in the context of thermodynamics, where energy can be given to (or taken from) a macroscopic system in two modes - as work and as heat. In this formula, $Q + W$ is the total energy flowing into the system in the form of heat and work. According to the principle of conservation of energy, this has to be equal to the increase in the total energy of the constituents making up the system. According to the interpretation of internal energy mentioned above, this is precisely what the left hand side of (8-7b) stands for.

While the formula (8-7b) holds for any process taking the system under consideration from an initial to a final state, both specified arbitrarily, one can consider a special case where the two states (say, 'a' and 'b') lie very close to each other in the state space, and the process under consideration involves only small amounts of energy exchange in the form of heat and work (we refer to such a process as a *small* one). For instance, the two states can define an infinitesimally small segment on a continuous path in state space, made up of a succession of equilibrium states, and corresponding to some quasi-static process involving the system. The formula (8-7b) can then be written in the form

$$\delta U = \tilde{\delta}Q + \tilde{\delta}W. \quad (8-7c)$$

The notation here needs an explanation because it relates to a subtle distinction of great relevance in thermodynamics. Since, as mentioned earlier, U is a thermodynamic variable, it has got well defined values in the states 'a' and 'b' lying close to each other in the thermodynamic state space, and δU denotes the small *change* in the value of U between 'a' and 'b'. On the other hand, $\tilde{\delta}Q$ and $\tilde{\delta}W$ do not refer to changes in the values of state functions, but rather, to small quantities relating to the *process* under consideration. One cannot talk of well defined values of 'work' and 'heat' for any given state, but only to amounts of work and heat involved in a *process*. This is why a different symbol is used for either of these quantities ($\tilde{\delta}Q$, $\tilde{\delta}W$) as compared with δU .

In the context of the quasi-static process, represented by a path in the state space, the small quantities occurring in (8-7c) are to be interpreted as going to zero in the sense of a limit, i.e., as *infinitesimally* small ones, in which case one uses the symbols $\tilde{d}Q$, $\tilde{d}W$, and dU respectively, where the former two are referred to as *inexact* differentials, in contrast to dU , which is an *exact* differential.

An inexact differential such as $\tilde{d}W$ cannot be expressed in the form $d\phi$ where ϕ is some state function (one having a well defined value for every point in the state space), and consequently, when it is summed up (i.e., *integrated*) over a succession of small segments lying on a path connecting two states, say, P, Q, the result is not determined solely by the two states P, Q, but also by the path connecting the two.

8.8.1 Equivalence of heat and work

One can see from equations (8-7a), (8-7b), that the unit of heat has to be the same as that of work or energy, namely, joule (J). This reflects the fact that the change in a process brought about by delivering some amount of heat to it (this corresponds to a process with the system kept in a diathermic enclosure) can be reproduced by means of an adiabatic change in the system where the only form of energy delivered to it is work. This is referred to as the principle of *equivalence* of heat and work. This means that 1 J of heat is that amount energy which produces the same effect as that due to the delivery of 1 J of work.

There is another commonly used unit of heat, defined in a somewhat different manner. The amount of heat necessary to raise the temperature of 1 g of water through 1 K is named 1 cal (*calorie*) of heat. Evidently, these two units have to be related in some way, since they pertain to the same physical quantity. Experimentally, it is seen that 1 cal is equivalent to 4.184 (or, approximately, 4.2) J. This is referred to as the *mechanical equivalent of heat*. Thus, the mechanical equivalent of heat is

$$\mathcal{J} = 4.184 \text{ J per cal.} \quad (8-8)$$

8.8.2 Work performed by a gas in a quasi-static process

Think of a fixed amount of gas in a cylinder fitted with a piston. Suppose that the area of cross-section of the cylinder, and of the piston, is α . Suppose that the pressure of the gas for some given position of the piston is p , i.e., the outward force exerted by the gas on the piston is $p\alpha$ (see fig. 8-5). We assume that the piston can move freely in the cylinder without friction at the walls. Then an equal inward force ($p\alpha$) has to be applied on the piston from outside so as to keep it in equilibrium in the above position. If now the external force on the piston be changed by a vanishingly small amount, then the piston will move very slowly.

Imagine such a process where the volume of the gas is made to change very slowly, a small change at an intermediate stage of the process being one in which the piston

moves by a small amount. In other words, at any intermediate stage, the gas is allowed to come to equilibrium with the volume of the gas measuring, say V , and pressure, say, p , when the external force on the piston is altered slightly as mentioned above, and equilibrium is allowed to be attained at the changed position of the piston.

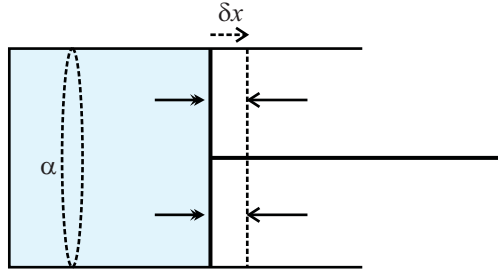


Figure 8-5: A small displacement of the piston during a quasi-static process involving a gas in a cylinder; double-headed arrows indicate force exerted by the gas on the piston while single-headed arrows depict the compensating force exerted from outside; the gas is in equilibrium at every stage of the process.

Let the small displacement of the piston in the intermediate stage referred to above be δx . For such a small displacement one can ignore the change in pressure brought about by this displacement and assume that the force exerted by the gas on the piston remains at the value $p\alpha$ during the displacement, as a result of which the work done by the gas on the piston amounts to $p\alpha\delta x$. In this expression, the product $\alpha\delta x$ measures the increase in volume of the gas (the term 'increase' being used here in the algebraic sense, where an actual decrease corresponds to a negative sign). Denoting this as δV , one concludes that the work done by the gas during a small displacement of the piston in a quasi-static process is $p\delta V$. This is the work done by the gas on some external agency, the agency that is made to exert the force on the piston so as to keep it in equilibrium at every stage during the process. Put differently, the work done *on* the gas by the external agency is

$$\tilde{\delta W} = -p\delta V. \quad (8-9)$$

As mentioned above, the symbol $\tilde{\delta W}$ is used here to denote a small amount of work,

and *not* to denote a small *increment* of some function, defined uniquely for the state of the system, in the process that can be assumed to be an infinitesimally small one. The volume, on the other hand, is a state function V , and δV denotes the change of this state function. Noting that the small changes considered here are actually *infinitesimally* small ones, δV represents an exact differential (dV), while $\tilde{\delta}W$ represents an inexact differential ($\tilde{d}W$).

One can now make use of the first law of thermodynamics, expressed in the form (8-7c), to obtain the following formula for the heat entering the system under consideration in an infinitesimal quasi-static process:

$$\tilde{\delta}Q = \delta U + p\delta V. \quad (8-10)$$

While we have considered a process (involving states arbitrarily close to each other) occurring in a gas, the above formula applies more generally to a pure *fluid*, for which state can be specified in terms of pressure and volume as independent thermodynamic variables.

If one now thinks of a finite quasi-static process represented by a path connecting two states, say, P and Q in the state space (see fig. 8-6 where two such paths are shown; the states 'a' and 'b' referred to above lie on one of these two paths), one has to imagine this to be divided into a large number of short segments and apply eq. (8-10) to each of these. One then has to add up the two sides of the resulting equations so as to obtain the mathematical formulation of the first law for such a process.

The formula one thereby arrives at is of the general form (8-7a), where the work done on the system (W) in the finite quasi-static process is given by the expression

$$W = - \sum_{(\text{path})} p\delta V. \quad (8-11)$$

In this expression one has to calculate the value of $p\delta V$ for each short segment of the

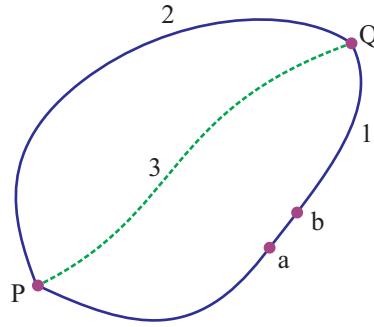


Figure 8-6: Two states represented by points 'a', 'b', lying close to each other in the state space, on a path marked 1 connecting points P and Q representing initial and final states of a quasi-static process; one other path, marked 2, between the same initial and final states is shown; the dotted line marked 3 represents a process that is not quasi-static and cannot be represented by a path; integral expressions for work and heat in terms of state variables cannot be obtained for such a process.

path under consideration and then has to add up all these values, which is indicated by the symbol $\sum_{(\text{path})}$. The result of the summation depends, in general, on the particular path under consideration. Strictly speaking, one has to assume that the segments are vanishingly small, i.e., the end points of each segment are infinitesimally close to each other, in which case the summation reduces to an *integration* along the path. In the end, one arrives at the following forms of the first law of thermodynamics.

$$\begin{aligned}
 Q &= \Delta U - W \\
 &= \Delta U + \sum_{(\text{path})} p \delta V \\
 &\rightarrow \Delta U + \int_{(\text{path})} p dV.
 \end{aligned} \tag{8-12}$$

The first of the three forms here is the *general* formula for the first law of thermodynamics which applies regardless of the nature of the system under consideration and of the type of process as well. The second and the third forms, on the other hand, are valid for any given amount of a pure fluid, and for a process where the fluid undergoes a quasi-static change, represented by a path in the state space. Of these, the second form is an intermediate one written as a reminder to indicate how the third form is arrived at.

Problem 8-2

Consider three states A, B, C of a gas represented in the p - V state diagram as in fig. 8-7, with the pressure and volume of each of the states A and B indicated in the figure. The dotted lines schematically represent adiabatic processes taking the gas from state B to C, and from A to C, in which the amounts of work done on the gas are W and W' respectively. Find the heat delivered to the gas in the quasi-static process taking it from A to B, represented by a straight line in the p - V diagram.

Answer to Problem 8-2

HINT: From the given data one has, in terms of notation as above, $U_C - U_B = W$, and $U_C - U_A = W'$; i.e., $U_B - U_A = W' - W$. By the first law of thermodynamics, this must be equal to $Q + W_{AB}$, where W_{AB} is the work done on the gas in the quasi-static process from A to B, and Q is the heat delivered to it. By eq. (8-11), W_{AB} is the area under the straight line from A to B in the p - V diagram, taken with a minus sign, i.e., $W_{AB} = -\frac{1}{2}(p_1 + p_2)(V_2 - V_1)$. Thus, finally, $Q = W' - W + \frac{1}{2}(p_1 + p_2)(V_2 - V_1)$

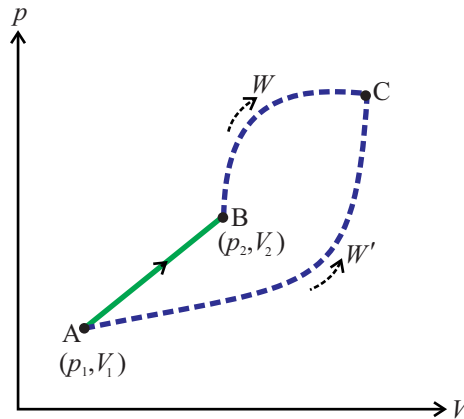


Figure 8-7: States A, B, C of a gas represented in the $p - V$ state diagram, with the pressure and volume of each of the states A and B indicated in parentheses; the dotted lines schematically represent adiabatic processes taking the gas from state B to C, and from A to C, in which the amounts of work done on the gas are W and W' respectively; considering a quasi-static process taking the gas from state A to B, represented by a straight line in the $p - V$ diagram, the heat delivered to the gas in this process can be obtained by invoking the first law of thermodynamics, as in problem 8-2.

8.8.3 Summary of the first law.

The zeroth law of thermodynamics helps us arrive at the concept of temperature, a state variable of great relevance for a thermodynamic system. The first law, in a similar spirit, leads us to the concept of internal energy, another state variable that can be interpreted as the total energy of the constituents of a system. Two other concepts of relevance are those of *work*, and *heat* which, however, are *not* state variables in that they do not have specific or well defined values for any given state of the system since they pertain to a *process* rather than to a state.

The concept of work is derived from mechanics. The amount of energy flowing into a thermodynamic system in the form of work is, in general, determined by referring to the agency that delivers the work. For instance, if work is performed on a fluid by means of rotation of a rod with attached vanes, the amount of work is obtained from the torque applied on the rod and the angle of rotation. The question that comes up here is, can the work be expressed in terms of thermodynamic variables relating to the system under consideration?

It is important to have a good idea as to what the question is about before one takes a look at the answer. The work delivered to a system can go to change its state of motion *as a whole*, like imparting a translation or rotation to the system. We are, however, interested not so much in the motion of the system as a whole, as in energy delivered to and distributed among the microscopic constituents of the system. Assuming that the entire energy flowing into the system in the form of work goes to its microscopic constituents, how can one express this energy in terms of the macroscopic state variables characterizing the system? A meaningful answer to this question is obtained only when the process of delivery of energy in the form of work is a *quasi-static* one.

This is where one comes up with the integral in the third expression in (8-12). For work delivered quasi-statically to a fixed amount of a fluid, the expression is $-\int_{(\text{path})} p dV$, which can be evaluated once one knows the pressure and volume at every intermediate state constituting the successive points on the path representing the process. Similar

expressions, involving additional thermodynamic variables, can be arrived at for quasi-static work delivered to a thermodynamic system of a more general description. No reference to the external agency delivering the work is required here.

Even as one expresses the work in terms of the thermodynamic variables pertaining to the system under consideration, one has to remember that the work depends on the *path* representing the quasi-static process in the state space. For two different paths in the state space (such as the paths marked 1 and 2 in fig. 8-6), the integral $-\int p dV$ will have, in general, two different values.

Thus, in the case of a quasi-static process, the work can be calculated in two different but equivalent approaches: one in the form of an integral expressed in terms of state variables pertaining to the system under consideration, and the other in terms of mechanical variables pertaining to the agency that delivers energy (in the algebraic sense) to the system in the form of work. In the case of a non-quasi-static process, on the other hand, the first of these two approaches is ruled out, and only way to quantitatively specify the amount of work is in terms of the external agency.

In contrast to the concept of work which is derived from mechanics, the concept of *heat* is a purely thermodynamic one. How to arrive at an expression for the amount of energy delivered as heat to a given system? The most general answer to this is the first law itself in the form of eq. (8-7a): knowing the change in internal energy *and* the work delivered to the system, one arrives at the expression for heat delivered. For a *quasi-static* process one can go further: one can derive an expression, similar to the above integral expression for work, involving state variables of the system under consideration.

One of the variables involved in such an integral expression is the *entropy* of the system. The concept of entropy is arrived at from the *second law of thermodynamics*.

Once again, the amount of energy flowing into a system in the form of heat in a quasi-static process depends on the path representing the latter. For instance, the heat delivered for the path marked '1' in fig. 8-6 differs, in general, from that for the path marked

‘2’.

As for internal energy, its change in any given process, whether quasi-static or not, is determined solely by the initial and final states, and not on the process connecting these two states.

As an application of all these ideas, consider the states depicted as P and Q in fig. 8-6. Two paths, marked ‘1’ and ‘2’, corresponding to quasi-static processes, are shown in this figure, while the dotted line, marked ‘3’, indicates a process between the same initial and final states that is not quasi-static and hence cannot be represented by a path in the state space. Denoting the energies delivered to the system in the form of heat and work in these three processes with suffices 1, 2, and 3 respectively, one has

$$\begin{aligned} U(Q) - U(P) &= Q_1 + W_1 \\ &= Q_2 + W_2 \\ &= Q_3 + W_3. \end{aligned} \tag{8-13}$$

Here Q_1 , Q_2 and W_1 , W_2 can be obtained from integral expressions involving state variables of the system, where the integrals are to be taken along the respective paths. On the other hand, Q_3 , W_3 cannot be expressed in the form of such integral expressions because the corresponding process cannot be represented by a path in state space. The important thing to note, however, is that, while Q_1 , Q_2 , Q_3 are, in general, all different, and so are W_1 , W_2 , W_3 , the pairwise sums $Q_1 + W_1$, $Q_2 + W_2$, $Q_3 + W_3$ are all equal, each representing the difference of internal energies of the system for the two states Q and P.

Problem 8-3

Imagine three states A, B, C, of a thermodynamic system and processes P_1 , P_2 , P_3 , taking the system from A to B, from B to C, and from C to A respectively. The amounts of energy given to the system in the form of heat and work in the process P_1 are respectively $Q_1 = 100$ J and $W_1 = 150$ J, the quantity of heat given in the process P_2 is $Q_2 = -50$ J (which means that heat is actually given out by the system), and the quantity of work done on the system in the process P_3 is $W_3 = 120$ J.

If the difference of internal energies of the system for the states B and C is $U_C - U_B = 20$ J, find the values of W_2 and Q_3 , where the symbols have their contextual meanings.

Answer to Problem 8-3

HINT: Since $Q_2 + W_2 = U_C - U_B$, one has, from the given data, $W_2 = 70$ J. Now, $U_B - U_A = Q_1 + W_1 = 250$ J, and hence $U_A - U_C = -(U_B - U_A) - (U_C - U_B) = -270$ J, which must be equal to $Q_3 + W_3$. Using the given value of W_3 , one gets $Q_3 = -390$ J, i.e., the system gives out 390 J of heat.

Problem 8-4

Imagine two states P, Q in a p - V state diagram connected by three paths, each corresponding to a quasi-static process, marked '1', '2', and '3', as shown in fig. 8-8. If the state P corresponds to pressure p_1 and volume V_1 , while Q corresponds to p_2 and V_2 , obtain the work done by the gas in each of the processes taking it from state P to state Q.

Answer to Problem 8-4

HINT: The work done *by* the gas in a quasi-static process is given by the expression $w = \int p dV$ (this differs by a negative sign compared to the work done *on* it; the integration is to be performed from the initial to the final state of the process), which is nothing but the *area* under the curve representing the process in the p - V diagram. Referring to fig. 8-8, the amounts of work for the three processes marked '1', '2', and '3' are respectively the areas under the straight lines AQ, PQ, and PB respectively, i.e., $w_1 = p_2(V_2 - V_1)$, $w_2 = \frac{1}{2}(p_1 + p_2)(V_2 - V_1)$, and $w_3 = p_1(V_2 - V_1)$.

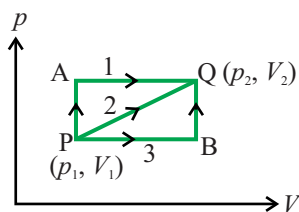


Figure 8-8: Two states, P and Q, in a p - V state space of a pure gas, connected by three paths marked '1', '2', and '3', corresponding to quasi-static processes; the path marked '1' consists of a constant-volume (*isochoric*) segment PA, and a constant-pressure (*isobaric*) segment AQ; the path '2' is a straight line segment connecting P and Q in which both the pressure and volume vary from point to point; and path '3' is made up of an isobaric segment PB and an isochoric segment BQ; the work done by the gas in the three cases turn out to be different, as in problem 8-4.

8.9 Intensive and extensive variables

The thermodynamic variables characterizing a system can be classified into two broad groups, namely the *intensive* and the *extensive* ones. The variables whose values are proportional to the quantity and volume of material making up the system are termed *extensive* ones. An example of an extensive variable is the internal energy of the system. If the mass of each macroscopic component making up a system as also its volume is doubled without changing its composition, the internal energy also gets doubled.

It is assumed here that, if after doubling, the system is divided into two equal parts, say, A and B, and the two parts are separated from one another, then each of the parts will be identical in all respects to the system one started with. This requirement is not met with if the *surface properties* of a system are taken into consideration. A more general definition, depending on the mathematical relationships among the various thermodynamical variables, then becomes necessary. We shall, however, not enter into such considerations here. I will indicate below another useful characterization of the extensive and intensive variables.

It is found, on the other hand, that the values of some thermodynamic variables remain unchanged upon a re-scaling of the system, i.e., a proportionate change in the size and quantity of material in the system. These are the *intensive* variables, examples of which are the pressure and temperature of the system.

The significance of extensive variables in thermodynamics differs from that of intensive ones. This can be explained by referring to an analogy from mechanics. Supposing that the displacement of a particle along any given direction is δx and the force acting on it in that direction is F , the expression for the work done in the displacement is known to be $F\delta x$. Now think of a small change in the thermodynamic state of a system. The heat supplied to the system or the work done on it in such a small change can be expressed in terms of one or more products of this form. Each product is, in other words, of the form $A\delta X$, where A and X stand for thermodynamic variables for the system under consideration.

referring to such a differential expression, the variables A and X featuring in it are seen to be, respectively, an intensive and an extensive variable for the system. This means that, in a thermodynamic change, the intensive variables play the role of *driving forces* and the extensive variables that of *generalized co-ordinates*. The changes in the extensive variables can then be interpreted as a generalized *displacements* under the action of the driving forces.

For instance, we have already seen that the work done by a fixed quantity of a gas is of the form $p\delta V$. Here the pressure, which is an intensive variable, can be looked upon as the driving force and the volume, an extensive variable, as the generalized co-ordinate whose change δV is analogous to a displacement.

8.10 The kinetic theory of gases

8.10.1 Macroscopic and microscopic descriptions

When a system is described from the point of view of thermodynamics, it is looked at as a *macroscopic* system. This means that the states of the system are described in terms of only a few variables relating to the system *as a whole*, without reference to possible states of its microscopic constituents. For instance, the pressure, volume, temperature, and internal energy of a gas are some of the variables pertaining to its thermodynamic, or macroscopic, state, and are termed the thermodynamic variables of the gas, not all of these being, however, mutually independent. In an equilibrium state (commonly referred to as a *state* in brief) of the system under consideration, all these variables which are defined and measured without direct reference to its microscopic constituents have definite values, which is why such a state is termed a *macrostate*.

In contrast, if one thinks of the microscopic constituents of the system (for our purpose the molecules will be taken as the microscopic constituents) then the states of motion of all these constituents taken together constitute a *microstate* of the system. Evidently, the description in terms of microstates is much more detailed and complex as compared to the description in terms of macrostates. However, such a microscopic description is,

in a sense, a fundamental one, and the question comes up as to how the microscopic and macroscopic descriptions of a system are related to each other.

Imagine that the states of motion of all the molecules of an isolated sample of gas are known at any given instant of time, and that the dynamical principles describing the evolution of the states of all these molecules with time are also known (for instance, the states of motion of the molecule can change by elastic collisions as in collisions between tiny hard spheres). Then starting from the given microstate, and invoking the given dynamical principles, can one arrive at a description of the macrostate of the gas? In general, the gas as a whole cannot be expected to be in a state of equilibrium. But can one, for instance, establish, from the microscopic mechanical considerations, that the gas, considered as a whole, will approach a state of equilibrium (depending on its volume and initial energy) deriving the correct relation between the macroscopic state variables in the equilibrium state, such as the pressure, volume, and temperature ? Such an approach of relating the microscopic and macroscopic descriptions of the states of a system is referred to as *kinetic theory*.

It is found that one can indeed arrive at meaningful results provided one makes a number of assumptions in relating the enormously large number of microscopic variables to only a few relevant macroscopic ones. The basic assumption is the one of *molecular chaos* where the instantaneous positions and velocities of the individual molecules are taken to vary in a *random* manner from one molecule to another. By an appropriate process of *averaging* over the microscopic positions and velocities of the individual molecules, and by appropriately interpreting the macroscopic variables in terms of these microscopic averages, a number of useful results are arrived at.

The approach of kinetic theory has proved useful in the case of gases. The macroscopic state of a pure gas under given conditions can be conveniently described in terms of the mole number (see sec. 8.10.2 below), volume, pressure, and temperature, while alternative descriptions in terms of other sets of variables are also possible. In the case of a gaseous mixture, one needs to specify the mole numbers of all the components in the mixture in order to arrive at a complete description.

I will now outline the above approach of kinetic theory in arriving at the the ideal gas *equation of state* (sections 8.10.3, 8.10.4, 8.10.5, 8.10.6), where a set of additional assumptions relating to an *ideal gas* are made use of.

8.10.2 Mole number

In the macroscopic description, the quantity of matter is indicated in terms of *mole number*. Consider a material made of a single kind of constituent units, say, a pure gas of molecular weight M . Then a quantity of the gas with mass equal to the molecular weight expressed in grams, (i.e., M g) corresponds to one mole. However, the material under consideration need not be a pure gas or a pure solid but may be any collection involving a single type of units, like, say, a number of ions in an electrolyte or a number of free electrons in a metal (see sec. 1.4.1). In cases such as these, the mole number (written as ‘mol’ in the SI system) refers to the number of elementary constituents in the collection, where 1 mol of the collection contains as many constituents as there are atoms in 12 g of carbon-12. This number is referred to as the *Avogadro number*, its value being

$$N_A = 6.02 \times 10^{23} \text{ mol}^{-1} \text{ (approx).} \quad (8-14)$$

Thus, if the number of constituents in a collection be N , then the corresponding mol number is given by

$$\nu = \frac{N}{N_A}. \quad (8-15)$$

In this equation, the unit of ν is mol, while N bears no unit, which explains the unit of the Avogadro number, as stated in eq. (8-14).

The reference to the Avogadro number serves as the link between the macroscopic and the microscopic descriptions of any given quantity of a substance. In spite of this, however, the mole is basically a unit to be used in the macroscopic description of a system. This means that there has to be an operational procedure for measuring the mole number without referring to the number of microscopic constituents. In any case, in the

thermodynamic description, the number of microscopic constituents is not a precisely defined number but is usually a quantity characterized by *fluctuations* as in any statistical specification. It is the *average* of the fluctuating quantity that corresponds to the mole number - a macroscopic variable.

8.10.3 The ideal gas

The volume of the vessel (say, V) containing a gas is usually very large compared to the space a single molecule of the gas occupies, where V gives the volume of the gas in the thermodynamic description. Indeed, the molecules are so small in extent that, for most purposes, these may be taken to be point masses with zero spatial extent. This is a convenient approximation in the macroscopic description. Moreover, the intermolecular forces being, in general, weak and of a short range, these forces can also be ignored for most purposes and the molecules may be assumed to be freely moving particles within the confines of the containing vessel. This also happens to be a very useful approximation.

An ideal gas is a hypothetical substance for which the molecules are point particles and the intermolecular forces are non-existent.

In the thermodynamic description, however, there is no place for reference to the molecular size or the intermolecular forces and the ideal gas is defined in a different way, by imposing the requirement that the mole number (ν), pressure (p), volume (V), and temperature (T) of the gas are to be related as

$$pV = \nu RT. \quad (8-16)$$

This is referred to as the *equation of state* of an ideal gas (or, in brief, the ideal gas equation), where R is a constant referred to as the *universal gas constant* (or the gas constant, in brief), its value being

$$R = 8.31 \text{ J} \cdot \text{mol}^{-1} \cdot \text{K}^{-1} \text{ (approx).} \quad (8-17)$$

Starting from the definition of an ideal gas according to the microscopic point of view and invoking the kinetic theory of gases, one can arrive at an equation describing the behavior of the gas that is consistent with eq. (8-16). In the following sections I will outline to you this approach of the *kinetic theory* of gases.

8.10.4 Pressure of a gas in kinetic theory

We will now see how the pressure of a gas (an ideal gas, that is,) is explained in the kinetic theory of gases and how it can be related to the *average* value of an expression relating to the state of motion of a typical molecule of a gas.

Suppose that the molecules of a gas are enclosed within a volume V . Referring to any given point O in this volume, imagine a small area around this point as shown in fig. 8-9 (marked ABCD in the figure), where this element need not be rectangular in shape. The molecules on either side of this area element are in incessant motion and keep on crossing this element either from the left to the right or from the right to the left (here 'left' and 'right' refer to the two sides of the area element with reference to the figure). Imagine a Cartesian co-ordinate system with the point O as the origin and with the x -axis chosen perpendicular to the area ABCD, as in the figure.

Think now of the x -component of the velocity of a typical molecule of the gas (we will refer to this as the x -velocity in brief). This x -velocity may have any value ranging from $-\infty$ to $+\infty$. However, for the sake of convenience we will assume that the x -velocity can have any one of a large number of discretely distributed values in the above range. In other words, the x -velocity will be assumed to be a discrete variable instead of a continuous one. Let u_i denote any one of these possible discretely distributed values of the x -velocity, where the suffix i can take up values $1, 2, 3, \dots$, corresponding to values, respectively, u_1, u_2, u_3, \dots of the x -velocity.

Imagine now the molecules of the gas to be grouped in accordance with the values of their x -velocities, and label the groups with indices $1, 2, 3, \dots$, the molecules of the i th group being the ones with u_i as their x -velocities. Considering any one of the molecules in this (i th) group, let the y - and z -velocities of this molecule be respectively v and w .

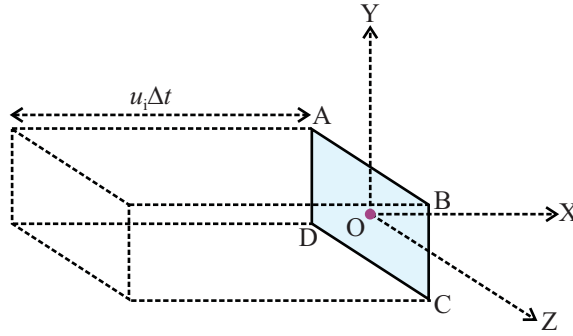


Figure 8-9: A small element of area ABCD imagined around any chosen point O in the volume occupied by a gas; molecules, in crossing this area carry momentum from one side to the other; considering a time interval δt , molecules with x-velocity u_i contained within a prism of length $u_i \Delta t$, with the area ABCD as base, will cross the area in this time interval, each molecule carrying with it an x-momentum mu_i , where u_i stands for a possible value of the x-component of the velocity.

Suppose that, at any given instant, this molecule crosses the area ABCD from the left to the right. In this case, if u_i happens to be negative, then the molecule will actually cross over from the right to the left and so, for the sake of concreteness, we assume u_i to be positive for the time being. Evidently, in crossing the area ABCD, the molecule will carry with it a certain amount of momentum whose x-component (we refer to this quantity as the x-momentum) will be mu_i where m is the mass of a molecule. Along with the x-momentum, some amount of y-momentum and z-momentum will also be carried by the molecule, but the *average* of the y-momentum or z-momentum carried through the area ABCD in any given interval of time will be zero.

Since the molecules of the gas move about with all possible velocities, and all the directions within the volume V are equivalent with respect to the molecules, there will, in all probability, be a molecule with velocity components $(u_i, -v, w)$ crossing the area ABCD in any given time interval corresponding to the one with components (u_i, v, w) referred to above. The total y-momentum carried by these two molecules crossing from the left to the right will then be zero. Considering in this manner all the possible y-velocities of the molecules, the total y-momentum (and, similarly, the z-momentum) carried through the area ABCD from the left to the right in any given time interval can be seen to be zero. Note that this logic, however, does not apply to the x-momentum

since the molecule with components (u_i, v, w) crosses the area from the left to the right whereas the one with components $(-u_i, v, w)$ crosses it from the right to the left.

Considering now a time interval δt , how many molecules with x-velocity u_i will cross the area element ABCD in this interval? Imagine a prism with the area element ABCD as base and of length $u_i \Delta t$, as shown in fig. 8-9. Evidently, all molecules having x-velocity u_i and lying within this cylinder will be able to cross this area element from the left to the right in the interval δt , each carrying with it an x-momentum mu_i . The number of such molecules is given by the expression $n_i \alpha u_i \delta t$ where n_i stands for the number of molecules per unit volume of the gas, with x-velocity u_i , and α denotes the area of the element ABCD.

The total x-momentum carried by these molecules in the time interval δt is then $(mn_i u_i^2) \alpha \delta t$, where n_i stands for the number of molecules per unit volume belonging to the i th group.

In other words, the *x-current* of the x-momentum (i.e., the x-momentum transported in the x-direction per unit area per unit time) due to these molecules is $mn_i u_i^2$. One would have obtained the *same* result if one considered a negative value of u_i instead of a positive one.

Check this out. You have to be careful about the signs of the various quantities involved (see below). Note that what we want to calculate here involves *two* directed quantities - the property that is being transported (the x-momentum) and the direction of transport normal to the y-z plane along the positive direction of the x-axis. It is actually a component of a *tensor* that is involved here.

The total x-current of the x-momentum is obtained by summing up the contributions of all the groups of molecules with various different x-velocities, and is given by $\sum_i mn_i u_i^2$, where the sum runs over all the values of the index i labeling the groups. This we now identify as the x-component of *force* (x-force in brief; rate of transfer of x-momentum) per unit area exerted through the area element ABCD (the normal to which is along the positive x-direction; let us agree to call it an x-area) by the part of the gas to the left of

the area on the part to the right. Let us provisionally denote this sum by the symbol p_x .

In continuation of what has been said above, it would seem that, in the summation $\sum_i mn_i u_i^2$, one is to include only those indices i for which u_i is positive. However, considering a group of molecules for which the x-velocity is negative, the number of molecules crossing from the positive to the negative side per unit time will be $-n_i u_i$ (check this out). Each of these molecules carry a x-momentum mu_i (which is actually a negative quantity). Thus the x-momentum carried by these molecules from the positive to the negative side per unit time is $-m_i n_i u_i^2$. An equivalent way of saying this is that there occurs a transfer of $m_i n_i u_i^2$ in the opposite direction, i.e., from the negative to the positive side, per unit time. In other words, one has to sum over *all* the groups of molecules, and not just over those with positive x-velocity, in order to arrive at the expression for rate of transfer of x-momentum in the positive x-direction.

Let the mean number of molecules per unit volume of the gas be n . Then, the probability of a molecule, chosen at random, to belong to the i th group will be $P_i = \frac{n_i}{n}$. In other words one can write

$$p_x = \sum_i mn P_i u_i^2. \quad (8-18a)$$

Here u_i^2 stands for the squared x-velocity of a molecule in the i th group, while P_i stands for the probability of the molecule belonging to that group. Then, in accordance with the basics of the theory of random variables (see sec. 8.24) the expression $\sum_i P_i u_i^2$ is nothing but the *mean* value (also referred to as the *expectation* value) of the squared x-velocity. Denoting the x-velocity of a typical molecule (which can belong to any of the groups mentioned above) by u , the squared x-velocity will be denoted by u^2 , the mean value of which we denote by $\overline{u^2}$. This implies

$$p_x = mn \overline{u^2}. \quad (8-18b)$$

As we have seen, the rates of transfer of y-momentum and z-momentum through the x-area are both zero. In a similar manner, one can talk of the y-force per unit area

through a y-area (for which the x-force and z-force would be both zero) and the z-force per unit area through a z-area (for which the x-force and y-force would be zero). These will be given, respectively, by the expressions

$$p_y = mn\overline{v^2}, \quad p_z = mn\overline{w^2}. \quad (8-18c)$$

We can now summarize the results arrived at. We consider, for the sake of simplicity, a small cubical volume element around any given point O in the region occupied by the gas, the edges of the cube being parallel to the three co-ordinate axes of any chosen co-ordinate system. The above results then imply that the force exerted per unit area by the portion of gas inside this volume on the adjacent portion of the gas through any of the faces of the cube is perpendicular to the face under consideration and directed away from the interior of the cube, being given by one of the expressions in equations (8-18b), (8-18c).

The important thing to note here is that all these expressions are, in reality, the *same*. This is because, all the three directions along the three co-ordinate axes (or, for that matter, any other direction) are equivalent with reference to the gas as a whole, since the molecules of the gas move about in all directions in an *uncorrelated* manner. As a consequence, even as the three components of the velocity of any particular molecule as also the squares of these components may differ from one another, the *mean values* calculated for all the molecules taken together will be the same. One can thus write

$$p_x = p_y = p_z = \frac{1}{3}mn\overline{c^2}, \quad (8-19a)$$

where

$$c^2 = (u^2 + v^2 + w^2), \quad (8-19b)$$

stands for the squared speed of a molecule.

The fact that the force exerted by a portion of the gas on an adjacent portion is directed along the normal to the surface of separation and is the same for all the orientations of

that surface can now be recognized as the principal characteristic of the stress force in a fluid and indicates that our derivation is consistent in this respect. The value of this force per unit area in any given direction is nothing but the bulk stress taken with a negative sign, or the *pressure* in the gas, for which we arrive at the expression

$$p = \frac{1}{3} m n \overline{c^2}. \quad (8-20)$$

We thus see that according to the kinetic theory, the pressure in an ideal gas is the result of *momentum transfer* from one region of the gas to an adjacent region through the common surface of separation. It is related to the *average value* of a quantity depending on the random motion of the molecules of the gas, namely to their mean squared velocity. Denoting the *square root* of the mean squared velocity as C , one can express the pressure in the form

$$p = \frac{1}{3} m n C^2. \quad (8-21a)$$

Here

$$C = \sqrt{\overline{c^2}} \quad (8-21b)$$

is referred to as the root mean squared (RMS) velocity (or speed) of the gas molecules.

With reference to the area ABCD in fig. 8-9, the x-momentum transported from the left to the right per unit time is equal and opposite to the rate of transport of x-momentum from the right to the left, i.e., the force exerted by one part of the gas on another is equal and opposite to the force exerted by the latter on the former.

8.10.5 The kinetic interpretation of temperature.

The kinetic theory of gases starts from the assumption that the molecules of a gas are in incessant random motion (also referred to as *thermal* motion) even when the gas is in a macroscopic equilibrium state. Considering any quantity relating to the dynamical state of a molecule, the value of that quantity will be found to vary from one molecule

to another but the *average* of all those values will be a well defined physical quantity relating to the macroscopic state of the gas. For instance, one can look at the kinetic energy of the molecules.

If the squared velocity of any molecule be denoted by c^2 then its kinetic energy will be $E = \frac{1}{2}mc^2$, where m is the mass of the molecule.

Here c^2 should not be confused with the squared velocity of light in vacuum, the latter being a common usage of the symbol.

The average kinetic energy of the molecules of the gas will then be

$$\overline{E} = \frac{1}{2}m\overline{c^2} = \frac{1}{2}mC^2. \quad (8-22)$$

where C stands for the RMS velocity of the molecules of the gas. Since, for an ideal gas, there is no interaction among the molecules the *potential energy* of the molecules can be taken to be zero. In other words, the expression (8-22) gives us the mean energy of the gas molecules.

Now think of two vessels with an ideal gas in each, and refer to the two gases as A and B. Assume that there is thermal contact between A and B so that these may exchange energy in the form of heat, but not in the form of work, including work relating to matter exchange (the quantity of gas in each vessel being, thus, fixed). The question that now comes up is, what determines the direction of heat flow here, i.e., whether heat will flow from A to B or from B to A? According to the principles of thermodynamics (including the *entropy principle*, refer to section 8.11) as also to empirical observations, heat will flow from the hotter to the colder gas. One can also look at the problem from the point of view of kinetic theory.

The basic process by which energy is transferred from one gas to another as heat, is that of *molecular collision* (see section 8.10.7). The collisions in the present instance do not occur directly between the molecules of the two gases but are indirect ones where the

molecules of each gas collide with those of the diathermic wall with which they are in contact. However, for the sake of simplicity, one may ignore the intermediary role of the diathermic wall and may imagine direct collisions to take place between the molecules of the two gases, still arriving at meaningful results. We can further assume that the collisions are perfectly *elastic* in nature, i.e., the total kinetic energy of two molecules is conserved in a collision between them. In other words, the kinetic energy lost by one molecule will be gained by the other.

1.

Collisions among molecules is not possible in an ideal gas since the latter is made up of point-like particles. One can, however, consider a gas made up of molecules of finite but arbitrarily small size so as to arrive at the concept of temperature in kinetic theory since all one needs for this purpose is the fact that the collisions between molecules do occur, while the *frequency* of collisions is not of relevance.

2.

In reality the collisions between molecules can be *inelastic* in nature where the total kinetic energy between molecules, instead of being conserved, is either gained or lost because of energy exchange with the internal vibrational and rotational motions of the molecules. If, however, the molecules are point-like particles without internal structure, or behave effectively like tiny hard spheres, the collisions are necessarily elastic.

Of the two molecules, say, 'a' and 'b', involved in a collision, which one is likely to lose energy to the other? Suppose for the sake of concreteness that, before the collision, 'a' has a higher energy compared to 'b'. Can one then say that 'a' will lose energy to 'b' in the collision? Looking at just one single collision, energy may, in fact, possibly flow the *other* way, i.e., from 'b' to 'a' (see sections 3.17.7.10, 3.17.7.11). If one were to observe the results of a *large* number of collisions, however, with the velocities of 'a' and 'b' prior to the collisions being randomly distributed, one would find that, *on the average*, energy is transferred from molecules of the gas with a higher mean energy to those of the other (the basic idea is out-

lined in section 3.17.7.11).

The criterion of energy transfer from one gas to another by molecular collisions can be stated in two ways: on the one hand, energy flows from the gas with a higher temperature to the one with a lower temperature; and on the other hand, it turns out that energy flow occurs from the gas with a higher mean molecular energy to the one with a lower mean energy (refer to section 3.17.7.11). This suggests that temperature should be related to the mean molecular energy. One can, indeed, define a scale of temperature based on such a relation between the two, one referred to as the *ideal gas scale* (or, in brief, the *gas scale*) of temperature. Imagine an ideal gas in an equilibrium state. The temperature (T) of the gas in that state in the gas scale is then a quantity proportional to the mean molecular kinetic energy ($\overline{E} = \frac{1}{2}mC^2$) of the gas:

$$\overline{E} = \alpha T, \quad (8-23)$$

where α is an appropriate constant. Choosing this constant to be $\frac{3}{2}k_B$ (k_B = Boltzmann's constant = $1.38 \times 10^{-38} \text{ J}\cdot\text{K}^{-1}$), the gas scale can be made to coincide with the *thermodynamic* scale of temperature introduced earlier (refer to section 8.4.3; see also sections 8.12, 8.15.4). One then has the kinetic theory interpretation of temperature for an ideal monatomic gas expressed in the following form

$$\overline{E} = \frac{1}{2}mC^2 = \frac{3}{2}k_B T. \quad (8-24)$$

It is to be mentioned here that, in the above derivations, \overline{E} stands for the mean *kinetic energy of translation* of the molecules of the gas. In addition to this, the molecules may possess kinetic energy of rotation and vibration that make their appearance for a gas made up of polyatomic molecules. For a monatomic gas, on the other hand, the kinetic energy of the molecule is the same as its kinetic energy of translation. In any case, the above formula refers to the kinetic energy of translation alone.

8.10.6 The ideal gas equation of state

We now return to equation (8-21a) which was deduced as the kinetic interpretation of pressure of an ideal gas. Combining this with eq. (8-24) one arrives at

$$p = nk_{\text{B}}T. \quad (8-25)$$

Here p denotes the pressure of the gas, T the temperature in the ideal gas scale (which we will assume to be equivalent to the absolute thermodynamic scale as also to the SI or Kelvin scale; see sections 8.4.3, 8.4.4, 8.15.4), n is the number of molecules per unit volume of the gas, and k_{B} denotes the Boltzmann constant. If the volume of the gas be V and the total number of molecules be N , then we have $n = \frac{N}{V}$. Again, the mole number ν is related to N by eq. (8-15), which thus gives

$$p = \frac{\nu N_{\text{A}} k_{\text{B}}}{V} T. \quad (8-26)$$

In this expression, if one replaces the product $N_{\text{A}} k_{\text{B}}$ with R , then the value of the constant R is found to agree with (8-17). This means that the constant R is nothing but the gas constant introduced earlier. One thereby arrives at an expression for pressure in conformity with the ideal gas equation of state (eq. (8-16)):

$$p = \nu \frac{RT}{V}. \quad (8-27)$$

At the same time, another important relation can also be worked out by multiplying both sides of eq. (8-24) with $N = \nu N_{\text{A}}$. This gives, on the left hand side, the expression $N\bar{E}$, i.e., the total kinetic energy of translation of the molecules of the gas. Since there is no force of interaction between the molecules of an ideal gas, their potential energy can be taken to be zero and hence the expression $N\bar{E}$ also gives, in the kinetic theory interpretation, the *internal energy* of the gas. Thus, eq. (8-24) leads to the following relation between macroscopic (or thermodynamic) variables pertaining to the gas

$$U = \frac{3}{2} \nu N_A kT = \frac{3}{2} \nu RT. \quad (8-28)$$

In arriving at the above expression for the internal energy, however, I have assumed implicitly that the latter derives solely from the kinetic energy of translation of the molecules of the gas under consideration. This assumption presupposes that, first, that there is no contribution to the internal energy coming from the energy of mutual interaction between the molecules (the ideal gas assumption) and, secondly, that there is no contribution coming from possible *internal modes* of the molecules. This latter assumption is valid to a good degree of approximation for a *monatomic* gas. In the case of a diatomic or a polyatomic gas, the expression for the internal energy gets modified because of contributions from the *rotational* and *vibrational* energies of the molecules.

The interesting thing to note in eq. (8-28) is that it has no place for the pressure or volume of the gas. In other words, *the internal energy of a given amount of an ideal gas at any given temperature does not depend on its volume or pressure*. This, indeed, can be taken as a defining characteristic of an ideal gas.

Any *real* gas, however, is found to deviate to some extent from the ideal behavior expressed by equations (8-16) and (8-28).

I repeat that eq. (8-28) holds only for a monatomic ideal gas. For a polyatomic ideal gas (i.e., one with no intermolecular interaction and with negligible molecular volume), the factor $\frac{3}{2}$ on the right hand side is to be replaced with some other appropriate factor depending on the molecular structure. However, the equation of state (8-27) continues to hold provided that, in the kinetic theory description, the molecules are of negligible volume and have negligibly small interaction between one another.

For instance, if one determines experimentally the values of the quantity $\frac{pV}{T}$ for a fixed quantity of any real gas at various temperatures and pressures, one finds that these values are not all the same. The deviation is found to be relatively more appreciable at

high pressures and low temperatures while, at low pressures and high temperatures, the deviation is seen to be negligible. In other words a real gas behaves like an ideal gas at low pressures and high temperatures. Unless the context demands otherwise, a real gas is commonly assumed to obey the ideal gas equations.

A number of corollaries follow from the equation of state (8-16) of an ideal gas.

1. *Boyle's law.* At a constant temperature, the volume of a fixed amount of a gas is inversely proportional to its pressure:

$$V \propto \frac{1}{p} \quad (\nu \text{ and } T \text{ constant}). \quad (8-29)$$

2. *Charles' law.* At a constant pressure, the volume of a fixed amount of a gas is proportional to its temperature:

$$V \propto T \quad (\nu \text{ and } p \text{ constant}). \quad (8-30)$$

3. *Law of pressures.* At a constant volume, the pressure of a fixed amount of a gas is proportional to its temperature:

$$p \propto T \quad (\nu \text{ and } V \text{ constant}). \quad (8-31)$$

4. *Avogadro's law.* At any given pressure, temperature and volume, all gases have the same number of molecules and mole number.

8.10.6.1 Ideal gas: isothermal and adiabatic processes

An isothermal process is one that occurs at a constant temperature. For a given quantity of an ideal gas, an isothermal process is characterized by a constant value of the product pV during the process (eq. (8-29)):

$$pV = A, \quad (8-32a)$$

where

$$A = \nu RT. \quad (8-32b)$$

Here ν and T stand for the mole number and the temperature of the gas under consideration.

Suppose now that ν mol of an ideal gas undergoes an isothermal process (recall from sec. 8.4.7 that an isothermal process has to be necessarily quasi-static in nature) at temperature T from a volume V_1 to V_2 . The work performed by the gas in the process can be obtained by noting that the work performed during an infinitesimal expansion from volume V by an amount δV is $p\delta V$ (refer to eq. (8-9) which gives the expression for the work done *on* the gas). Adding up such expressions for the entire process, which corresponds to an integration from volume V_1 to V_2 , one obtains the required expression for the work done as

$$W = \int_{V_1}^{V_2} p dV. \quad (8-33)$$

Making use of equations (8-32a) and (8-32b), this reduces to

$$W = \nu RT \ln \left(\frac{V_2}{V_1} \right). \quad (8-34)$$

Considering, on the other hand, a quasi-static *adiabatic* process, the work done by the gas in such a process from volume V_1 to V_2 is obtained from eq. (8-33) by making use of eq. (8-102) in sec. 8.21.4:

$$pV^\gamma = B \text{ (say)}, \quad (8-35a)$$

where γ denotes the ratio of the specific heat at constant pressure and the specific heat at constant volume of the gas (see sections 8.21.3 and 8.21.4), and can be assumed to be a constant for the gas in the present context.

The expression for the work done in the quasi-static adiabatic process then works out

to

$$W = \frac{B}{\gamma - 1} (V_1^{1-\gamma} - V_2^{1-\gamma}). \quad (8-35b)$$

Finally, making use of the equation of state (eq. (8-16)) of the gas, this can be expressed in the form

$$W = \frac{R(T_1 - T_2)}{\gamma - 1}, \quad (8-35c)$$

where T_1 and T_2 stand for the initial and final temperatures of the gas (check this out).

Problem 8-5

Imagine 1 mol of an ideal monatomic gas to be at a pressure p and volume V , and suppose that it is made to undergo a free expansion (see sec. 8.11 below) up to a volume $2V$ without any heat exchange with any other body. What is the internal energy of the gas after the expansion? If now a heat $Q = \frac{1}{2}pV$ be given to the gas at constant volume, what will be its pressure? What will be its pressure if the gas is now compressed isothermally back to volume V ? Find the work done on the gas and the heat given out by it in this isothermal process.

Answer to Problem 8-5

HINT: The temperature in the initial state is $T = \frac{pV}{R}$, and hence the internal energy is $U = \frac{3}{2}RT = \frac{3}{2}pV$. Since, in the free expansion process, the gas does not perform any work, and moreover, does not exchange heat with any other system, its internal energy remains unchanged at U . On addition of Q quantity of heat at constant volume (no work done by the gas), its internal energy gets changed to $U' = U + Q = 2pV = 2RT$, i.e., the temperature rises to $T' = \frac{4}{3}T$ (reason this out) and hence its pressure will be $p' = \frac{RT'}{V'} = \frac{2}{3}p$ (since $V' = 2V$). As the gas is now compressed isothermally back to volume V , its pressure becomes $\frac{p'V'}{V} = \frac{4}{3}p$. The work done on the gas in this process will be $W = RT' \ln(\frac{2V}{V}) = (\frac{4}{3} \ln 2)pV$ (refer to eq. (8-34)). Since the internal energy remains unchanged in an isothermal process, the heat given out by the gas will also be $Q' = (\frac{4}{3} \ln 2)pV$.

Problem 8-6

Consider two vessels of volumes V and $V' = 2V$ containing a gas at pressures p and $p' = \frac{p}{4}$ and at temperatures T and $T' = \frac{3T}{2}$ respectively. The two vessels are connected by a tube with a valve attached to it, the valve being initially closed. If the valve is now opened so as to bring about an equalization of the pressures, with the temperatures kept unchanged at their initial values, find the common pressure p'' in the vessels, assuming the gas to be an ideal one.

Answer to Problem 8-6

HINT: If the vessels initially contain respectively ν and ν' mol of the gas, then one has, $pV = \nu RT$ and $p'V' = \nu' RT'$, i.e., $pV = 3\nu' RT$, or, in other words, $\nu = 3\nu'$. The total quantity of the gas in the two vessels taken together is thus $\frac{4}{3}\nu$ mol. The final pressure p'' established after the opening of the valves is obtained from the relations $p''V = \nu'' RT$, and $p''V' = (\frac{4}{3}\nu - \nu'') RT'$, where ν'' stands for the mol number of the gas in the first vessel after the equalization of pressure. This gives $\nu'' = \frac{4}{7}\nu$, and thus $\frac{p''}{p} = \frac{\nu''}{\nu}$, i.e., $p'' = \frac{4}{7}p$.

8.10.7 Random motion of molecules: Molecular collisions.

The concepts of macroscopic and microscopic states of a system have already been introduced. The equation of state (8-16) tells us that, knowing any three of the four quantities p , T , V , and ν one can specify a macrostate (or a thermodynamic state, often referred to as, simply, a state) of an ideal gas.

Apart from these variables, a number of other variables like the internal energy and entropy (a state variable whose definition depends on the *second law of thermodynamics*; refer to sections 8.12, 8.13 below) can also be made use of in specifying a macrostate of the gas.

Knowing a macrostate, however, does not mean that the states of motion of all the molecules of the gas are known. It is a knowledge of the latter that enables one to specify a microstate. Such knowledge is, in reality, a near impossibility. Indeed, the state of motion of even a *single* molecule keeps on changing in course of time in an almost unpredictable manner. This is why any attempt to refer to the values of microscopically

defined quantities has to be abandoned in favor of a description in terms of *averages* of such quantities.

If one tries to follow the motion of any particular molecule of the gas, one will find that it is incessantly changing its speed and direction of motion in a random manner owing to collisions with other molecules and with those of the walls of the containing vessel. The motions of the molecules are, moreover, *uncorrelated* with one another.

If the gas molecules are assumed to be point particles then the probability of collisions among the molecules becomes negligibly small. If, in addition, the molecules are assumed to exert no force on one another then this would mean that the molecules would fly about within the volume occupied by the gas independently of one another. In reality, however, the molecules do not conform to these idealized assumptions. One can instead assume that the molecules are like tiny hard spheres and that their interaction is limited to a strong repulsion on contact, much like billiard balls hitting one another. The hard sphere collisions can, moreover, be assumed to be *elastic* in nature. One thereby arrives at a more realistic description of the molecules and their collisions than the one based on the ideal gas assumptions. At the same time, a number of meaningful results and conclusions are obtained relating to the behavior of the gas.

Apart from the collisions among the molecules of the gas, there occur collisions of the gas molecules with the molecules of the walls of the containing vessel. Such collisions are relevant in the context of energy and material exchange between the gas and the surroundings. These can also be taken to be elastic collisions, causing abrupt changes in the magnitude and direction of the velocity of a molecule on impact.

As a result of collisions, a molecule exchanges energy and momentum with other molecules, and there takes place a *transport* of a number of dynamical quantities like energy and momentum from one region within the gas to other regions. The kinetic theory of gases allows one to calculate transport coefficients like the *thermal conductivity* and *coefficient of viscosity* of the gas. In this, one needs the concept of the *mean free path* of the

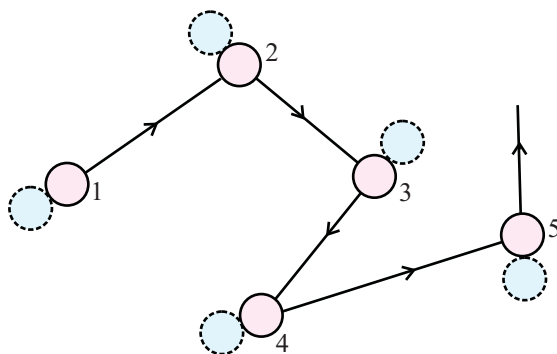


Figure 8-10: The path followed by a single molecule of a gas suffering collisions with other molecules; the positions of the molecule at successive collisions are marked 1,2,...; the molecules with which these collisions take place are depicted with dashed circles; in between successive collisions, the molecule moves in straight line paths; the average of all these straight line segments is the mean free path.

molecules.

8.10.7.1 Mean free path

Fig. 8-10 depicts schematically the path of a molecule undergoing a number of successive collisions with other molecules of the gas. In between any two successive collisions, the molecule follows a straight line path since there is no interaction between molecules except at the time of impact. The path followed by the molecule appears to be a random one and the successive straight line segments, termed free paths of the molecule, vary greatly in length from one another. But looking at a large number of successive collisions, one can arrive at an *average* length of these free paths. This average is the same for all the molecules of the gas since the motion of one molecule does not differ in its statistical features from another, and is referred to as the *mean free path* of the molecules.

For a given gas, the mean free path turns out to be proportional to the temperature and inversely proportional to the pressure of the gas. The constant of proportionality depends on the range of interaction (the minimum distance of approach between two molecules in an impact under the present assumptions) among the molecules, referred to as the *collision diameter*.

8.10.8 Brownian motion

Transport processes in a gas like the conduction of heat caused by the transfer of kinetic energy and viscous drag caused by the transfer of momentum, are all related to the random molecular motions, punctuated by molecular collisions. Another common transport process is *diffusion* caused by the transfer of the molecules themselves from regions of higher number density (i.e., the number of molecules per unit volume) to ones of relatively lower number density.

Analogous to the random motion of the gas molecules, particles suspended in a gas or a liquid also exhibit a random motion referred to as the *Brownian motion*, where such a motion can provide one with a visual demonstration of how the random motions occur at a microscopic level. For instance, *colloid* particles in a liquid move about by Brownian motion. A higher number density of these particles in one region of the liquid compared to other regions results in a diffusive process, driven by Brownian motion, whereby the densities are equalized.

The Brownian motion of a particle suspended in a fluid is caused by *unbalanced impact* of the molecules of the fluid on the particle at various instants of time. This is illustrated schematically in fig. 8-11 where the fluid molecules (small dashed circles) surrounding a Brownian particle (large solid circle) in two successive positions are shown. In the first position, the impact of molecule A causes the particle to fly to the second position where the impact by B sends it off to a new position. There being no correlation between successive impacts, the resulting motion of the particle is a random one.

The Brownian particle has to be large compared to the molecules in order that its motion may be visible to the naked eye or with the help of a microscope. However, it should not be too large for the unbalanced force to be averaged out before the particle can move by an appreciable distance.

Because of the random nature of the Brownian motion, the mean displacement of a Brownian particle in any interval of time of sufficient length has to be zero. However, the

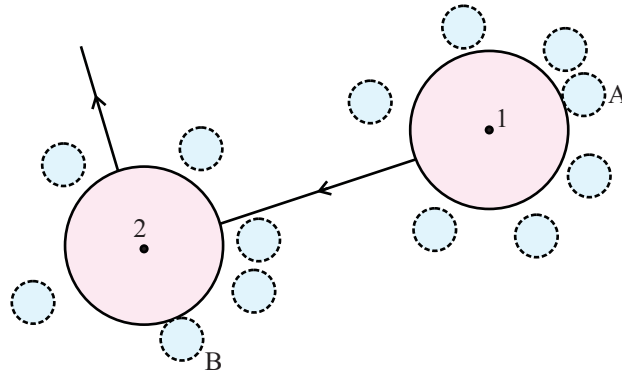


Figure 8-11: Brownian motion of a particle suspended in a fluid, by unbalanced impact of the fluid molecules; two positions of the particle, marked 1 and 2, are shown; in the first position an unbalanced force acts on the particle due to impact by the molecule A, while in the second position a similar force arises due to impact by B.

mean of the *squared* displacements between successive impacts with the fluid molecules is found to be proportional to the time interval of observation. The mean squared displacement per unit time increases with the *temperature* of the fluid. With increasing temperature, the thermal motion of the fluid molecules becomes more vigorous (for instance, we have seen that the mean kinetic energy of the molecules is proportional to the temperature) and hence the average momentum received by the particle in an impact also increases, thereby causing the root mean squared displacement per unit time to increase.

8.10.9 Mean speed and most probable speed

If one considers any quantity relating to the motion of individual molecules, its value at any given instant will be found to differ for all the different molecules of the gas. For instance, the speed of a molecule ($c = \sqrt{u^2 + v^2 + w^2}$, u , v , w being respectively the x-velocity, y-velocity, and z-velocity of the molecule) may have any value from 0 to infinity.

There occur two types of statistical variation in the value of any given dynamical quantity, say, μ (pronounced 'mu'), of a molecule. On the one hand, the values of μ at any given instant of time for the various different molecules may differ from one another. And, on the other, the value of μ for any given molecule may keep on changing randomly with time as the molecule suffers successive collisions. In defining the average

value of μ , one therefore can proceed in either of two ways: one can consider the average over the different values of μ for all the different molecules at any given instant of time, or else one can take the average over the values of μ at successive instants of time for one single molecule. Interestingly, both the approaches yield the same result when the system under consideration is in a state of thermodynamic equilibrium, and it is immaterial as to which average one is talking of.

However, a very high or a very low value of the speed is less likely compared to other, more moderate, values. The *probability* or likelihood for various values of the speed can be known from *Maxwell's velocity distribution formula*, which I will state for you in sec. 8.10.10 below. A conclusion that can be drawn from this formula tells us that the *most probable speed*, or the speed most likely to occur among all the possible values from zero to infinity is given by the expression

$$\tilde{c} = \sqrt{\frac{2kT}{m}} = \sqrt{\frac{2RT}{M}}, \quad (8-36)$$

where m stands for the mass of a molecule while M denotes the molecular weight of the gas expressed in kg (also referred to as the molar mass).

Maxwell's velocity distribution formula also gives the *mean speed* of the molecules as

$$\bar{c} = \sqrt{\frac{8kT}{\pi m}} = \sqrt{\frac{8RT}{\pi M}}, \quad (8-37)$$

Notice that the mean speed is not the same as the root mean squared speed, where the latter is seen from eq. (8-24) to be given by

$$C = \sqrt{\frac{3kT}{m}} = \sqrt{\frac{3RT}{M}}. \quad (8-38)$$

With an increase in the temperature, the molecules undergo more vigorous random motion and the most probable speed, the mean speed, and the root mean squared speed all increase in proportion to \sqrt{T} .

Each of the above expressions can be expressed in terms of the pressure and density of

the gas as well. Making use of the equation of state of the gas one obtains $\frac{RT}{M} = \frac{pV}{\nu M}$. Here the product of the mole number (ν) and the gram molecular weight (M) is nothing but the mass of the gas, and hence $\frac{\nu M}{V}$ is its *density* (say, ρ). One can then express the quantities \tilde{c} , \bar{c} and C in terms of P and ρ . For instance, the expression for the RMS speed works out to

$$C = \sqrt{\frac{3P}{\rho}}. \quad (8-39)$$

Fig. 8-12 depicts schematically the probability distribution for the molecular speeds in a gas for any given temperature T . Here the probability density $p(c)$ for speed c has the following significance: considering a small range of speed from, say, c to $c + \delta c$, the fraction of molecules having their speeds in this range is given by $p(c)\delta c$. One observes that, in the graph of $p(c)$ against c , the most probable speed corresponds to the maximum value of $p(c)$, as the definition of \tilde{c} implies. At a higher temperature T' ($> T$) the entire graph shifts towards the right as shown by the dotted curve in fig. 8-12. This corresponds to a more vigorous thermal motion of the molecules of the gas.

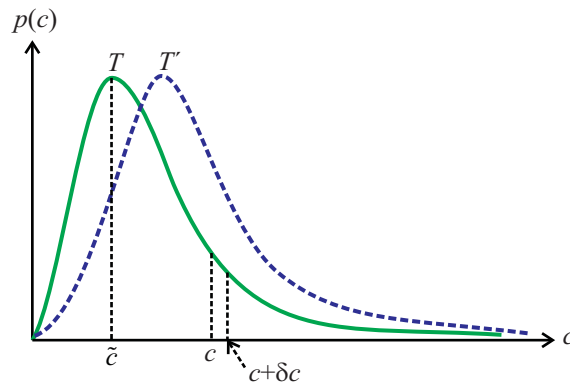


Figure 8-12: Speed distribution function $p(c)$ as deduced from Maxwell's velocity distribution formula, for any given temperature T (schematic); the area under the curve for any small range c to $c + \delta c$ gives the fraction of molecules with their speeds in this range; for a higher temperature T' , the graph shifts towards the right, and the most probable speed \tilde{c} increases.

Problem 8-7

The molar mass of helium is $4.0 \times 10^{-3} \text{ kg} \cdot \text{mol}^{-1}$. Find the root mean squared speed of helium

molecules at 300 K. At what temperature will the root mean squared speed be (a) half, and (b) twice this value?

Answer to Problem 8-7

HINT: In the second expression for C in (8-38), put $M = 4.0 \times 10^{-3} \text{ kg}\cdot\text{mol}^{-1}$, $R = 8.31 \text{ J}\cdot\text{K}^{-1}\cdot\text{mol}^{-1}$, and $T = 300 \text{ K}$ to obtain $C = 1.37 \times 10^3 \text{ m}\cdot\text{s}^{-1}$. Since, for a given gas, $C \propto \sqrt{T}$, the root mean squared speed gets halved at $T' = \frac{T}{4} = 75 \text{ K}$, and it gets doubled at $T'' = 4T = 1200 \text{ K}$.

8.10.10 Maxwell's velocity distribution formula

The kinetic theory tells us that the molecules of a gas possess velocities that vary randomly from one molecule to another. Thus, considering an ideal gas at equilibrium at any given temperature T (measured in the ideal gas scale, which is defined so as to be consistent with the thermodynamic scale to be discussed in sec. 8.15.4), the velocity components (say, v_x, v_y, v_z), with reference to any chosen Cartesian co-ordinate system, of a molecule is in the nature of a *random variable*. As outlined in sec. 8.24, a random variable is characterized by a *probability distribution* over a set of possible values. The possible values of any one of the velocity components (v_x, v_y, v_z) range continuously from $-\infty$ to $+\infty$, where *relativistic* considerations are not taken into account.

The *special theory of relativity* (refer to chapter 17) imposes the restriction that the speed of a particle cannot exceed c , the *speed of light in vacuum*. This constrains each of the components v_x, v_y, v_z to lie in the range $-c$ to $+c$. Maxwell's velocity distribution formula (8-40) stated below gets altered when this restriction is taken into account. The modified formula is referred to as the *Maxwell-Jüttner* distribution formula. However, the consequences derived from this modified formula coincide, to a good degree of approximation, with those from the Maxwell formula unless the temperature T is extremely high such as that within the body of a star.

Since the possible values of each of the velocity components are distributed continuously, it is not meaningful to look for the probability of a molecule having precisely

specified values, say, v_x, v_y, v_z (which would be zero regardless of the values specified), while, on the other hand, it does make sense to ask as to what the probability would be for the components to lie within small ranges, say, within v_x and $v_x + \delta v_x$, v_y and $v_y + \delta v_y$, and v_z and $v_z + \delta v_z$. For sufficiently small intervals $\delta v_x, \delta v_y, \delta v_z$, this probability would be of the form $f(v_x, v_y, v_z)\delta v_x\delta v_y\delta v_z$, where the object of interest is the function $f(v_x, v_y, v_z)$, referred to as *Maxwell's velocity distribution function*. Once specified, it gives all relevant information relating to *average* values of arbitrarily specified functions of the velocity components such as the mean speed of and mean squared speeds of the molecules considered in sec. 8.10.9. I will now state for you the all-important formula for the velocity distribution function:

$$f(v_x, v_y, v_z) = \left(\frac{m}{2\pi k_B T}\right)^{\frac{3}{2}} \exp\left(-\frac{m(v_x^2 + v_y^2 + v_z^2)}{2k_B T}\right). \quad (8-40)$$

In this formula, m stands for the mass of each molecule (we assume all the molecules to be of the same mass), and k_B for the Boltzmann constant.

The velocity distribution function of eq. 8-40 is properly *normalized* in the sense that the probability of a molecule having velocity components that may lie *anywhere* within the range over all possible values, which is the integral of the distribution function taken over the entire interval from $-\infty$ to $+\infty$ for each velocity component works out to unity:

$$\int_{v_x=-\infty}^{+\infty} \int_{v_y=-\infty}^{+\infty} \int_{v_z=-\infty}^{+\infty} f(v_x, v_y, v_z) dv_x dv_y dv_z = 1. \quad (8-41)$$

As a corollary, the distribution function for any one of the three components, say for v_x works out to

$$\tilde{f}(v_x) = \left(\frac{m}{2\pi k_B T}\right)^{\frac{1}{2}} \exp\left(-\frac{mv_x^2}{2k_B T}\right). \quad (8-42)$$

where $\tilde{f}(v_x)$ has the interpretation that the probability of v_x lying within a small range v_x to $v_x + \delta v_x$, with the other two components *unspecified* (i.e., with their values lying anywhere within their respective ranges), is $\tilde{f}(v_x)\delta v_x$.

8.10.11 Partial pressure

Consider a closed chamber of volume V containing a mixture of two different ideal gases (say, A and B) at a temperature T , there being ν_1 and ν_2 moles of A and B in the chamber. The total number of moles is then $\nu = \nu_1 + \nu_2$, and it is this total mole number that determines the pressure at any point in the gas mixture.

More precisely, following the logic of the derivation in sec. 8.10.4, the total momentum transfer per unit area per unit time across a surface oriented in any specified direction around any chosen point is $\frac{1}{3}(m_1 n_1 C_1^2 + m_2 n_2 C_2^2)$, where the meanings of the symbols are as explained earlier, with suffixes '1' and '2' referring to the two gases in the mixture. On substituting $n_1 = \frac{\nu_1 N_A}{V}$, $n_2 = \frac{\nu_2 N_A}{V}$, one obtains the formula (8-43) below.

In other words, the pressure p at any point in the chamber is given by (see formula (8-27))

$$p = \nu \frac{RT}{V} = (\nu_1 + \nu_2) \frac{RT}{V}. \quad (8-43)$$

Imagine now a situation in which the chamber contains just ν_1 mol the component A at the temperature T , i.e., the same temperature as above. The pressure in the chamber in this situation is referred to as the *partial pressure* (p_A) of the component A in the previous situation where the chamber contained the mixture. Similarly, the partial pressure (p_B) of B in the mixture is defined to be the pressure in the chamber in a situation in which it contains just ν_2 moles of B at temperature T . Making use once again of the expression (8-27) for the pressure of an ideal gas one has,

$$p_A = \nu_1 \frac{RT}{V}, \quad p_B = \nu_2 \frac{RT}{V}, \quad (8-44)$$

and hence,

$$p = p_A + p_B. \quad (8-45)$$

The above definition of partial pressure can be extended in an obvious manner to a

gaseous mixture made of more than two components, and eq. (8-45) can be generalized to the statement that *the pressure of a gaseous mixture equals the sum of the partial pressures of all the components in it.*

Problem 8-8

The pressure in a vessel of volume $V = 1.0 \text{ m}^3$, containing a mixture of two gases A and B, is $p = 1.5 \times 10^4 \text{ Pa}$, at a temperature $T = 300 \text{ K}$. If the amount of A in the vessel is $\nu_A = 2.5 \text{ mol}$, find the partial pressure of B in the chamber.

Answer to Problem 8-8

HINT: The partial pressure of A in the chamber is $p_A = \nu_A \frac{RT}{V} = \frac{2.5 \times 8.31 \times 300}{1.0} \text{ Pa}$, i.e., $6.23 \times 10^3 \text{ Pa}$. Hence the partial pressure of B is $p_B = p - p_A = (15.0 - 6.23) \times 10^3 \text{ Pa}$, i.e., $8.77 \times 10^3 \text{ Pa}$.

8.11 Reversible and irreversible processes

Suppose two bodies A and B at two different temperatures T_1 and T_2 are kept in thermal contact, and that A is the hotter of the two bodies ($T_1 > T_2$). One observes in this situation that there occurs a flow of heat from A to B and their temperatures are equalized as the bodies reach a state of equilibrium. In other words, *heat flows from the hotter to the colder of two bodies kept in thermal contact.* The reverse process of heat flowing from a colder to a hotter body is never found to occur spontaneously. In other words, the flow of heat from a hotter to a colder body is an *irreversible process*.

Indeed, every spontaneous process in nature has a definite *direction* associated with it, and it cannot proceed in a reverse direction all by itself. It is thus a general characteristic of natural processes that they are all irreversible.

Another instance of an irreversible process is the free expansion of a gas. Suppose that a closed chamber is divided into two halves with a removable partition inserted in it, and a gas is kept in one of these two as in fig. 8-13(A), the other half being vacant. If now the partition is removed then the gas will undergo a free expansion into the vacant

half and will eventually occupy the entire region in the chamber with uniform density as in fig 8-13(B). This is an irreversible process because the gas can never be found to compress itself spontaneously into one of the two halves. Here the chamber is assumed to be isolated from its surroundings so that it is not affected by any external system by way of exchange of heat or work.

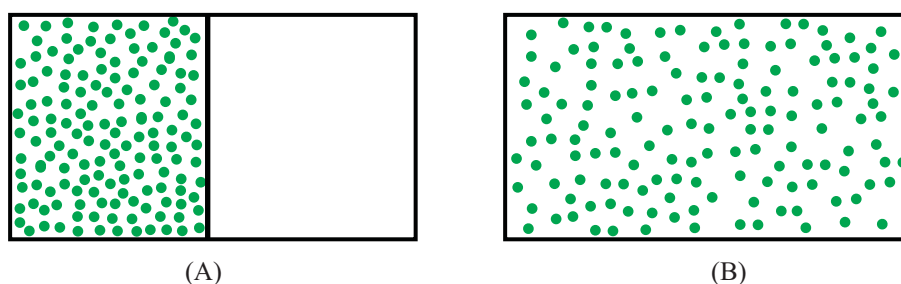


Figure 8-13: Free expansion of a gas; in (A), the gas is contained in the left half of the chamber, being separated from the vacant part on the right by a removable partition while in (B) it is uniformly distributed in the entire chamber as a result of a free expansion.

While an irreversible process cannot spontaneously proceed in a reverse direction, the question that one can ask is, whether it can be made to occur in the reverse direction with the aid of external systems so as to completely annul the original change. As an example, think of the flow of heat from a hotter body A to a colder body B. If the composite system made up of A and B is assumed to be isolated from the rest of the world then there does not occur a change of state in any external system as a result of the spontaneous flow of heat from A to B when the two are kept in thermal contact. If then the reverse process of heat flow from B to A is made to occur with the help of a third system (say, S) then, at the end of the reversed process, this third system has to return to the state it started from, i.e., in other words, the change in the state of S has to be a *cyclic* one since, otherwise, the original process of heat flow from A to B cannot be said to have been truly reversed.

In practice, the process of transfer of heat from a colder to a hotter body is made to occur with the help of a refrigerator, where usually the atmosphere plays the role of the hotter body, while the coolant used in the refrigerator (a mixture of a volatile liquid and

its vapor) plays the role of the third, auxiliary, body. The coolant is made to repeatedly pass through a cyclic process, and heat is transferred from a colder to the hotter body in each cycle. However, in order to make the cyclic process occur, some amount of *work is to be performed* on the coolant from an external source during each cycle. Consequently, the amount of heat flowing into the hotter body is larger than the amount extracted from the colder one. Evidently, this cannot be considered to be an exact reversal of the original process of flow of heat from the hotter to the colder body. This corroborates the statement that the flow of heat from a hotter to a colder body is an irreversible process.

However, a process can conceivably be made to occur by special arrangement such that the corresponding reverse process can be brought about without any residual change remaining in any of the participating systems. In other words, after the completion of the direct and the reversed processes, all the systems taking part in the processes return to their initial states. Such processes are referred to as *reversible* ones. In reality, however, such a reversible process can be made to occur only in a limiting sense and no real process can be completely reversible. This means that the concept of a reversible process is an idealized one. Still, one can adopt measures in making a process occur such that it can be described as a reversible process to a good degree of approximation.

As an example, one can think of the process of expansion of a gas kept in a cylinder fitted with a movable piston. As I have mentioned above, the free expansion of a gas is an irreversible process. However, if a gas is allowed to expand very slowly by a quasi-static process (or, more precisely, by a process that is close to being a quasi-static one) and if the friction between the piston and the wall of the cylinder is made vanishingly small, then the expansion of the gas will be close to a reversible process.

Referring to fig. 8-14, let A be the initial position of the piston and B its position at the end of the expansion. Let, at any intermediate stage of the expansion process, the pressure of the gas be p , i.e., the outward thrust exerted by the gas on the piston is $p\alpha$, where α stands for the area of cross-section of the piston. In accordance with the condition stated above, let the force of friction resisting the movement of the piston be assumed to be zero.

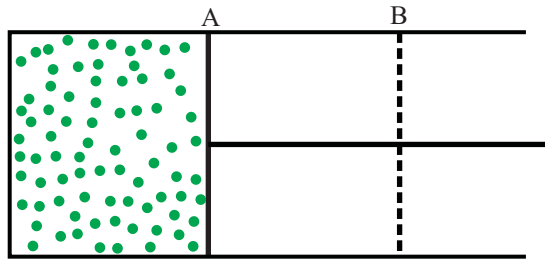


Figure 8-14: Expansion of a gas kept in a cylinder fitted with a piston; the gas is made to expand quasi-statically from the position A of the piston up to the position B; friction between the piston and the walls of the cylinder is assumed to be negligible; the process can be reversed with little residual change left anywhere.

If an equal and opposite thrust be exerted on the piston from an external source at every stage of the process so that it cancels the thrust $p\alpha$ exerted by the gas and if, in the initial position A of the piston, it be given a vanishingly small outward velocity, then it will keep on moving with this small velocity and the gas will expand outward very slowly. This can be described as a quasi-static process to a good degree of approximation. Suppose that the slow expansion is allowed to continue till the piston reaches the position B.

Suppose now that, as the piston reaches the position B, a vanishingly small inward velocity is imposed on it so that now the gas starts contracting slowly till it reaches the position A, when the piston is made to stop where, at every intermediate point of time, an external force is applied on the piston, as in the first leg of the process, so as to balance the thrust exerted by the gas at that instant. Evidently, at the end of the expansion and contraction processes taken together, the gas returns to its initial state. Can there be any residual change left anywhere else? Assuming the velocity of the piston to be vanishingly small at every instant during expansion and contraction, the external system coupled to the piston does not perform any net work on the gas. The only work done arises from the need to impart the small initial velocity to the piston, stopping it at the position B, giving it a small velocity once again, and finally stopping it at position A. However, since the velocities involved are vanishingly small, these changes are also negligible.

Thus, the process of expansion of the gas has been reversed in the process of con-

traction. What prevents the process from being made to occur in this ideal manner is the force of friction that cannot be eliminated completely, and the small but non-zero amount of work done by the external system(s) coupled to the piston, where both these factors lead to *dissipation* of energy in the participating systems. In other words, the concept of a reversible process is an idealization, though such a process is not ruled out *in principle*. It is a process that can be made to occur in a limiting sense. Any process occurring in practice is irreversible to some extent.

That every process occurring in practice has a direction associated with it, is an empirically observed feature of diverse natural phenomena and of processes made to occur by various man-made set-ups. Indeed, this constitutes the empirical basis of the *second law of thermodynamics*. The second law can be formulated in several alternative ways, but all of those relate to this directed nature of processes involving thermodynamic systems.

While the zeroth law of thermodynamics leads to the concept of temperature as a thermodynamic variable, and the first law to the one of internal energy, the second law of thermodynamics leads to the concept of yet another thermodynamic variable of great significance, namely, the *entropy*. It is the entropy that makes possible a precise formulation and interpretation of the second law of thermodynamics in the form of what is commonly referred to as the *entropy principle*: *the entropy of an isolated system increases in all natural processes*; it can remain constant in a reversible process, which is an idealized one, but can never decrease.

8.12 Entropy: thermodynamic definition

Consider a quasi-static process taking a thermodynamic system from a state A to some other state B, as depicted schematically in fig. 8-15 where the thermodynamic state space of the system is assumed to be a two dimensional one for the sake of simplicity (for instance, the system under consideration may be some definite quantity of an ideal gas). The states are represented by two points in the state space, while the process is depicted by a path connecting the two points.

Let P be any intermediate state in the process, for which the temperature is T . Considering a small segment of the path around P, let us assume that, during the traversal of this small segment, the system receives $\tilde{\delta}Q$ amount of heat and perform $\tilde{\delta}W$ amount of work. As indicated in sec. 8.8, these correspond to *inexact differentials*.

In the case of a gas, while $\tilde{\delta}W$ corresponds to an inexact differential and, when summed up over a large number of small segment of the path extending from A to B in the figure, does not result in a value that depends only on the initial and final states A, B, the infinitesimal quantity $\frac{\tilde{\delta}W}{p}$ *does* correspond to an exact differential, since it is nothing but δV , the change in volume where the latter is a state function. On summing over, this simply gives $V_B - V_A$, the difference in volumes for the states A, B. Here $\frac{1}{p}$ acts as an *integrating factor* for the inexact differential $\tilde{\delta}W$.

However, consider now the small quantity $\frac{\tilde{\delta}Q}{T}$. Interestingly, *this* quantity does correspond to an exact differential, i.e., on being summed up over the small segments of the path from A to B, it yields a value that depends solely on the two states A and B. Taking A as a *reference state* fixed by convention, the value of the integral (the sum over small quantities $\frac{\tilde{\delta}Q}{T}$ in the limit when all these small quantities go to zero) depends on the state B alone, and hence defines a thermodynamic state function, which we define as the *entropy* of the state B.

In other words, the entropy of a state, where we now denote the latter by the symbol P, is defined as

$$S_P = \int_R^P \frac{\tilde{\delta}Q}{T}, \quad (8-46)$$

where the integration is performed over any arbitrary path in the thermodynamic state space connecting the reference state R, fixed by convention, with the state P in question.

Defined this way, the entropy depends additively on an undefined constant value since, changing over to a new reference state, say, R' , the entropy gets altered by the addition of the entropy of the state R, now defined with reference to R' .

1. One can then say, in analogy to the inexact differential $\tilde{d}W$, that the expression $\frac{1}{T}$ plays the role of an integrating factor for $\tilde{d}Q$.
2. The temperature T appearing in formula (8-46) is to be taken in the *thermodynamic* scale introduced in sec. 8.4.3 where, however, the scale was not defined in precise terms. In the present context, one may define the temperature in the thermodynamic scale as being the same as the temperature occurring in the formula (8-27). Even though the latter, referred to as the temperature in the absolute gas scale, is defined with reference to an ideal gas, it can nevertheless be made use of in arriving at the thermodynamic scale of temperature because the thermodynamic scale is one that does not actually depend on the properties of any specific system. You will find further considerations regarding the absolute scale of temperature in sec. 8.15.4. A practical realization of the absolute scale, to a good degree of approximation, is the SI scale of temperature, introduced in sec. 8.4.4.

A corollary to the above definition can be stated as *the principle of additivity* of entropy: in a system made up of two or more parts (i.e., subsystems) the change in entropy in any given process is the sum-total of changes in the entropies of the constituent parts. This, in turn, implies that the entropy of a system is an *extensive* thermodynamic variable.

As another useful corollary, one arrives at the following expression for the change in entropy of a gas in an infinitesimally small part of a quasi-static process:

$$dS = \frac{\tilde{d}Q}{T} = \frac{dU}{T} + p \frac{dV}{T}, \quad (8-47)$$

where the first law of thermodynamics, in the form of eq. (8-10), has been made use of.

Problem 8-9

The expression for internal energy of one mole of an ideal monatomic gas is $U = \frac{3}{2}RT$. Find the change in the entropy of ν moles of the gas in a quasi-static process in which its pressure and volume are both doubled.

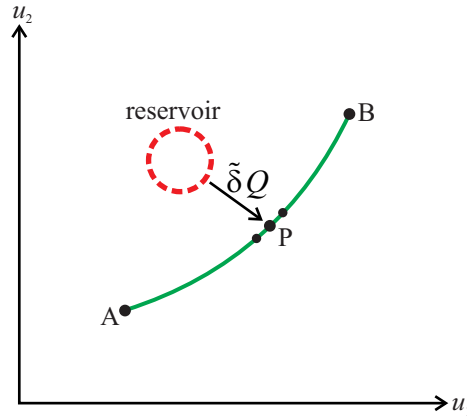


Figure 8-15: Illustrating the thermodynamic definition of entropy; A and B are points representing two states of a system in the thermodynamic state space, assumed to be two dimensional for the sake of simplicity (u_1 , u_2 are the the thermodynamic variables defining the axes in the state diagram); a path connecting these two states is considered, corresponding to a quasi-static process connecting the two states; for any point P on this path, if T denotes the temperature, in the thermodynamic scale, of the system, and δQ the heat given to it quasi-statically in an infinitesimal process represented by a small segment of the path around P (this may be the heat delivered from a heat reservoir, shown schematically by the dotted circle, whose temperature differs from T by an infinitesimally small amount), then the ratio $\frac{\delta Q}{T}$ corresponds to an exact differential and gives the change in entropy in the infinitesimal process; the sum of expressions of this form over all the segments covering the path from A to B gives the entropy of the state B relative to A; the entropy of any arbitrarily chosen state such as P, relative to any fixed reference state R, is given by an expression of the form 8-46.

Answer to Problem 8-9

HINT: The ideal gas equation of state, eq. (8-16), implies that the temperature increases to four times its initial value in the process under consideration. Denoting the thermodynamic variables in the initial and final states by suffixes '1' and '2', making use of the formula (8-47), and integrating over the infinitesimal change in entropy from the initial to the final state, the change in entropy in the process under consideration is seen to be

$$\Delta S = \nu \left(\frac{3}{2} R \ln \frac{T_2}{T_1} + R \ln \frac{V_2}{V_1} \right).$$

Substituting $\frac{T_2}{T_1} = 4$ and $\frac{V_2}{V_1} = 2$, one obtains $\Delta S = 4\nu R \ln 2$. Here the principle of additivity of entropy has been made use of, which implies that the entropy of ν moles of the gas is ν times the entropy for one mole, under identical conditions of pressure and temperature.

8.13 The second law of thermodynamics

With the entropy defined as in sec. 8.12, the second law of thermodynamics can be formulated as the *entropy principle* stated in sec 8.11: the entropy of an isolated system cannot decrease as a result of any process taking place within it (i.e., one in which no external system is involved); it always increases unless the process is a reversible one.

The isolated system under consideration may be made up of sub-systems and the process referred to above is one where the states of these subsystems are changed with reference to their initial states.

As already mentioned, a reversible process is an idealized one such that a second process can be made to occur in the reversed direction taking the system under consideration back to its initial state, which means that all the sub-systems are brought back to the states they started from. This requires that the process be a quasi-static one, occurring under infinitesimally small 'driving forces' (such as the temperature difference between two sub-systems between which a heat exchange takes place), and that 'dissipative' processes (such as frictional ones) do not take place. In the end, the necessary and sufficient condition for a process in a closed system to be reversible is that the change in its entropy be zero.

Suppose a closed system A is made up of the sub-systems B and C and suppose, in a process occurring within A, the changes in entropy of A, B, and C are respectively ΔS_A , ΔS_B , and ΔS_C . Then the second law of thermodynamics, together with the principle of additivity of entropy, implies that

$$\Delta S_A = \Delta S_B + \Delta S_C \geq 0, \quad (8-48)$$

where the equality sign corresponds to the ideal case of a reversible process.

The second law of thermodynamics can be stated in alternative ways as well. For instance, consider a process where a certain amount of work, say, W is performed on an

adiabatically enclosed system A, causing its internal energy to increase by W . One can then bring the system in thermal contact with a second system B, whereby this same amount of energy (i.e., W) is transferred to B in the form of heat ($Q = W$) while A returns to its initial state. In other words, an amount of work can be completely converted to heat given to a body. However, a second process *cannot* be made to occur where the heat is completely converted back to work, because that would constitute a violation of the entropy principle.

Another equivalent statement of the second law of thermodynamics is the following: heat can be extracted from a hotter body and completely transferred to a colder body, but the reverse process of extraction of heat from a colder body and complete transfer to a hotter body cannot occur.

Thermodynamics is a subject of vast proportions where a great number of consequences of practical relevance follow from a few basic principles (namely, the laws of thermodynamics). However, I will not pursue the subject further, and will consider only a few consequences in a very limited number of contexts relating to *heat engines* (sec. 8.15). However, before that, I include in sec. 8.14 below a few words on *statistical physics*, where the entropy principle is seen in a new light.

8.14 Statistical physics: Boltzmann's formula

Consider, for the sake of concreteness, a fixed quantity of a gas kept in a cylinder of a given volume and possessing a given amount of internal energy, where the gas is isolated from its surroundings. Though this corresponds to a unique equilibrium state of the system as a whole that we have referred to as a macrostate, it does not imply a definite *microstate* of the gas, where the latter involves a specification of the states of motion of each of its microscopic constituent. Referring to the approach of kinetic theory, for instance, while each molecule does have a well defined position and velocity at any given instant, there is a random variation of position and velocity *among* the molecules and, any other assignment of positions and velocities with the same total energy of all the molecules taken together would correspond to the same macrostate.

Thus, in reality, a macrostate corresponds to a large number of microstates so that, given a macrostate, one can talk of the microstates only in *statistical* terms, i.e., in terms of their *probabilities* (see sec. 8.24 for a brief introduction to random variables and probabilities).

As another instance, consider a fixed amount of gas in a cylinder of given volume, where now the gas, instead of being isolated from its surroundings, is kept in thermal contact with a heat reservoir (see sec. 8.4.6) at any given temperature, say, T . Thus, the equilibrium state (i.e., the macrostate) of the gas is the one corresponding to the given values of V and T .

In the case of the gas with a given volume isolated from its surroundings, the macrostate corresponds to given values of V and U , the latter being the internal energy of the gas. However, these values of V and U correspond to some definite value of T as well. Thus, the two ways to define the macrostate of the gas (one in terms of V and U , and the other in terms of V and T) do not differ fundamentally from each other. This is basically because of the fact that the gas is a macroscopic system, made up of a large number of microscopic constituents.

As mentioned above, this macrostate corresponds to a large number of possible microstates of the system, with some definite probability associated with each of these microstates. The total energies of the molecules in these microstates may differ from one another in spite of the macrostate being characterized by some definite value of the internal energy U , since U is now nothing more than an *average* of the energies of the possible microstates.

For a given macrostate of the system in thermal contact with a heat reservoir at temperature T (measured in the absolute scale; refer to sec. 8.15.4), what is the probability of occurrence of a microstate of energy E ? The answer is given by a fundamental formula in statistical physics, known as the *Boltzmann formula*:

$$p(E) \propto e^{-\frac{E}{k_B T}}, \quad (8-49a)$$

where k_B stands for the Boltzmann constant. The constant of proportionality can be worked out from the requirement that all the various probabilities have to add up to unity, and is written as Z^{-1} , where Z is termed the *partition function* of the macrostate under consideration. Thus, the Boltzmann formula finally looks like

$$p(E) = \frac{1}{Z} e^{-\frac{E}{k_B T}}. \quad (8-49b)$$

What is remarkable about this formula is that *it applies equally well to small systems in thermal contact with a heat bath*. Thus, consider a gas in thermal contact with a heat bath, where there is no interaction among the molecules of the gas, as a result of which the molecules move about independently of one another. If the gas is kept in thermal contact with a reservoir at temperature T , then each of its molecules interacts independently with the reservoir. At equilibrium, the probabilities of the microstates of the gas as a whole follow the Boltzmann formula (8-49b) and, at the same time, the possible states of *each molecule* also follow a formula of the same form, where now E and Z stand for a possible energy value of the molecule and the partition function for the single molecule respectively.

As an application of this fundamental formula of statistical physics, it is possible to derive from it *Maxwell's velocity distribution formula* for an ideal gas stated in sec. 8.10.10.

8.14.1 The statistical interpretation of entropy

Another result of great importance in statistical physics relates to an interpretation of *entropy* from a fundamental point of view. For this, let us refer once again to a fixed quantity of a gas in a cylinder isolated from its surroundings, so that its macrostate is specified in terms of the volume V and the internal energy U . Recall that this macrostate corresponds to a large number of microstates of the gas, where each microstate corresponds to some definite state of motion of each microscopic constituent of the gas.

Since we do not actually know what these microstates are, the number of possible microstates (call it W) for the given macrostate is a measure of our ignorance of the

microscopic details relating to the macrostate under consideration, or of the degree of *disorder* characterizing the macrostate because, the larger the value of W , the greater is the number of the unknown microscopic configurations underlying the state of the gas.

When the macrostate is specified in terms of V and U rather than V and T , the Boltzmann formula (8-49b) does not apply any more. Instead, all microstates with energy U turn out to be *equally probable*. However, for a system made up of a large number of microscopic constituents, the two turn out to have equivalent consequences, to a good degree of approximation.

Boltzmann made the fundamental contribution of relating the entropy of a system to this number W of the microstates corresponding to a given macrostate. More precisely, he proposed the following formula for the entropy (S) of a system:

$$S = k_B \ln W. \quad (8-50)$$

With this statistical interpretation of entropy, the entropy principle can be stated in a new light: in any natural process involving a number of systems isolated from all other systems, the total disorder of all the systems participating in the process increases and the disorder attains its maximum value as all the systems taken together reach the state of equilibrium. This is at times referred to as the *principle of disorder*.

Note that the Boltzmann formula (8-50) gives the entropy of a system *in absolute terms* without any arbitrary additive constant, which is apparent conflict with the definition (8-46) where the entropy for any state P of a system is defined up to an additive constant, depending on the reference state R.

Indeed, the Boltzmann formula includes additional information not contained in the principles of thermodynamics considered in the above sections, since it relates the macroscopic state of a system to its microstates while thermodynamic principles do not make overt reference to the microscopic constituents of the system. It is due to this feature of thermodynamic principles that one needs an *additional* thermodynamic

principle in order to remove the arbitrariness (relating to an additive constant) in the definition of entropy and to make it consistent with the Boltzmann formula. This additional principle is the *third law of thermodynamics* (or the *Nernst heat theorem* as it is commonly referred to) which states, in effect, that the entropy of any system goes to zero as the absolute zero of temperature is approached. This, at the same time, gives us the statistical significance of the absolute zero, namely it is a temperature at which only one single microstate is possible for any system (i.e., in formula (8-50), $W = 1$, giving $S = 0$). In turn, this is consistent with *quantum mechanical* principles - ones supposed to govern the behavior of all systems, whether microscopic or macroscopic.

We have seen that an alternative way of specifying a macrostate is in terms of the volume V and the temperature T of the system without overt reference to the energy U . In this alternative description of a thermodynamic state, the basic principle pertaining to the relation between a macrostate and the underlying collection of microstates is given by (8-49b) rather than the principle of equal probabilities mentioned above. Formula (8-50) is then no longer applicable in defining the entropy in microscopic terms, since one now needs to define first another thermodynamic parameter, namely, the *free energy* (F) of the system, and then introduce the entropy by means of the formula

$$F = U - TS, \quad (8-51)$$

where, as mentioned before, U stands for the average of the energies of the underlying microstates. For large (or macroscopic) systems for which the principles of thermodynamics are applicable, the two approaches (one where U, V are adopted as the basic variables, and the other where the basic variables are V, T) lead to the same values of all the relevant thermodynamic variables despite differences in how they are introduced into the theory.

8.15 Heat engines

A *heat engine* is a device to extract energy in the form of heat from a thermal reservoir at an appropriately high temperature (commonly referred to as a heat source; more

than one such heat sources may also be involved) and to transfer it to another body (commonly, a *prime mover*) in the form of work. Recall that heat exchange is a process of energy transfer between bodies that occurs by virtue of a temperature difference, while work is the energy given to a body or received from it when it is maintained within an adiabatic enclosure.

In the process of extracting heat and delivering work, the engine has to give up some amount of heat to a thermal reservoir at an appropriately low temperature (a heat sink; once again, more than one such heat sinks may be involved).

8.15.1 Explaining the idea of a heat engine

The material in a heat engine that exchanges energy in the form of heat and work is referred to as the *working substance* of the engine. The engine is operated by repeatedly making the working substance go through a *cyclic process*, where each cycle consists of a series of thermodynamic changes at the end of which the working substance returns to the state it started from.

As the working substance goes through a cycle, it extracts, say, Q_1 amount of heat from the heat source, and delivers W amount of work to the prime mover or to any other mechanical system, while at the same time giving up Q_2 amount of heat to the heat sink. Thus the net result of the cycle is to convert internal energy drawn in the form of heat (of net amount $Q_1 - Q_2$) into mechanical energy of the prime mover.

The fact that the engine cannot simply extract a certain amount of heat from a heat source and deliver the whole of it as mechanical energy to the prime mover, but has to give up a certain amount of energy as heat to the sink (which is a wasteful process from the point of view of energy conversion), is due the fundamental stricture imposed by nature on any and every thermodynamic process in virtue of the *second law of thermodynamics*, one of the several alternative forms of which is the *entropy principle* (refer to sections 8.11, 8.13, and 8.14).

The thermodynamic processes undergone by the working substance during a cycle of the

engine are, in general, irreversible ones and cannot be described graphically in a thermodynamic state diagram. However, a good way to understand and analyze the working of the engine is to imagine that the processes are made to occur quasi-statically (see sec. 8.3.3) so that the working substance passes through a succession of equilibrium states. The cycle can then be represented in the thermodynamic state diagram by a closed path. This closed path is made up of a number of segments, where each segment corresponds to some definite process constituting a stage of the cycle. For instance, the extraction of energy from the heat source constitutes a stage while the delivery of work to the prime mover corresponds to some other stage(s) in the cycle.

The working of a heat engine is conveniently described by means of such an idealized representation of the processes occurring in each cycle of the engine, wherein the cycle is assumed to be a *reversible* one. The results derived with reference to such an idealized representation turn out to be of considerable relevance for actual engines, in the sense of yielding *limiting values* to the performance characteristics of the latter.

In addition to the requirement that the processes involved in the working of the engine be made to occur quasi-statically, another closely related requirement for the processes to qualify as reversible is that all *dissipative* processes be absent, a dissipative process being one where the energy transferred to the microscopic constituents of a system cannot be retrieved from the latter. In order that a process may occur without dissipation, it has to be a quasi-static one.

In the following we will look at the working principles of a number of heat engines by making use of this idealization, assuming the processes constituting an operating cycle of the engine to be reversible ones.

8.15.2 The efficiency of a heat engine

In a cycle of operation of the heat engine, the working substance receives energy in the form of heat from one or more heat sources, the amount of heat received being

Q_1 , according to the notation of sec. 8.15.1. Similarly, the amount of heat rejected to the heat sink(s) is Q_2 , and the net work delivered to the prime mover is W (say). The processes of energy exchange in a complete cycle of the engine are shown schematically in fig. 8-16.

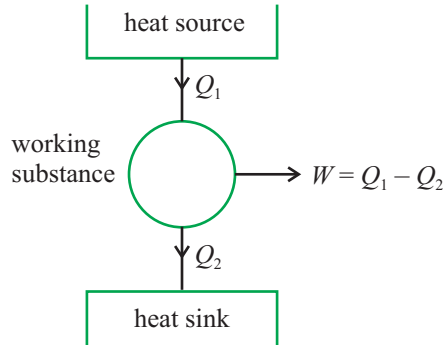


Figure 8-16: Energy exchange during a cycle of a heat engine; an amount of heat Q_1 is absorbed by the working substance from one or more heat sources, a part of which (Q_2) is given up to one or more heat sinks at lower temperatures, the rest ($W = Q_1 - Q_2$) being delivered as work to an external body, thereby increasing its mechanical energy; in a Carnot cycle, the extraction and the rejection of heat occur isothermally, at constant temperatures, say, T_1 and T_2 .

The principle of conservation of energy (or, equivalently in the present context, the first law of thermodynamics) tells us that the net energy taken in as heat has to be the same as the energy delivered as work, i.e.,

$$W = Q_1 - Q_2. \quad (8-52)$$

From the point of view of energy conversion, Q_2 represents a wastage since it causes a diminution in the amount of mechanical energy delivered to the prime mover from the value Q_1 , the energy extracted from the heat source. Thus, a basic target in the development of a heat engine is to make Q_2 as small as possible in comparison with Q_1 , where the limit to which this can be accomplished is set by the second law of thermodynamics.

Thus, an important performance characteristic of a heat engine relates to the ratio

$$\eta = \frac{W}{Q_1} = 1 - \frac{Q_2}{Q_1}, \quad (8-53)$$

referred to as the *efficiency* of the engine. A decrease in the ‘wastage’ (decreed by the second law) of energy Q_2 , relative to Q_1 implies an increase in the efficiency since a relatively greater fraction of Q_1 is then converted to mechanical energy of the prime mover.

In the theoretical analysis of heat engines, one works out the ‘ideal’ efficiency by assuming all the thermodynamic processes making up a cycle of the engine to be reversible ones. The efficiency of energy conversion of an actual engine based on these processes will be less than the ideal efficiency since there occurs energy dissipation, an wastage of energy *in addition to* the one arising due to the limitation set by the second law of thermodynamics. This energy dissipation is caused by such effects as friction between moving parts of the engine, viscous forces in the fluids used in the engine, enhanced energy dissipation due to turbulence in the fluids, and thermal conduction and radiation processes. I have to mention, however, that the reduction of efficiency due to the occurrence of dissipative processes is *also* decreed by the second law itself. One can minimize this reduction to a negligible value, but the restriction that the efficiency can still not attain the value unity cannot be done away with.

In the following, when we speak of the efficiency of an engine, we mean the ideal efficiency where all dissipation effects are ignored.

Even when each single process in a complete cycle of the engine is operated quasi-statically and dissipation effects in each single process are eliminated in a limiting sense, the overall cycle has to conform to the second law, which is why the efficiency of the engine has to be less than unity, an efficiency of unity being possible, again, only in a limiting sense. In the context of the overall cycle, the fact that $q_2 > 0$ (and hence $\eta < 1$), is a restriction decreed by the second law, and the limiting situation $Q_2 = 0$ can be achieved only

as an idealization. This explains the apparent paradox that the efficiency of a reversible engine, for which the entropy change of all the participating systems considered together turns out to be zero in a complete cycle, is, generally speaking, less than unity. Put differently, the second law implies that the entropy of any closed system increases when dissipative processes occur in it, but it does not demand that the efficiency of a heat engine has to be unity in the absence of dissipative processes. To repeat, the second law does impose the restriction that the entire heat drawn from a source cannot be converted to work (except in a limiting sense), and a part of the heat has to be rejected to a sink which implies that the efficiency has to be less than unity, though it can approach the value unity in a limiting sense.

8.15.3 An ideal heat engine: the Carnot cycle

Imagine an ideal heat engine in which the working substance extracts heat from a *single* heat source at a constant temperature, say T_1 , and likewise gives up part of this energy as heat to a single heat sink at a constant temperature T_2 ($T_2 < T_1$). The remaining stages in a complete cycle of the engine are then necessarily made up of two quasi-static adiabatic processes where the working substance exchanges energy in the form of work with external systems. Exchange of energy in the form of work may also take place during the isothermal processes of heat uptake and heat rejection. The net work delivered by the working substance is given by the expression (8-52), in terms of which the efficiency is obtained from (8-53).

The complete cycle then consists of four stages, of which two correspond to isothermal processes, and two to adiabatic ones. Such a cyclic process is termed a *Carnot cycle*, and an engine in which the working substance is made to follow the Carnot cycle is referred to as a Carnot engine.

For instance, a Carnot engine can be made to operate with an *ideal gas* as the working substance. A Carnot cycle can also be made to run with a *real gas* or, for that matter, with some other working substance since the expression for the efficiency (eq. (8-54b))

remains the same regardless of what the working substance is. The ideal gas Carnot cycle consists of (a) an isothermal expansion at temperature T_1 , (b) an adiabatic expansion, cooling the gas from T_1 to T_2 , (c) an isothermal compression at temperature T_2 , and (d) an adiabatic compression heating up the gas from T_2 to T_1 . The complete cycle can be represented on a p - V state diagram as in fig. 8-17, where the gas undergoes the above processes as it passes from state A to state B, from B to C, from C to D, and from D back to A respectively.

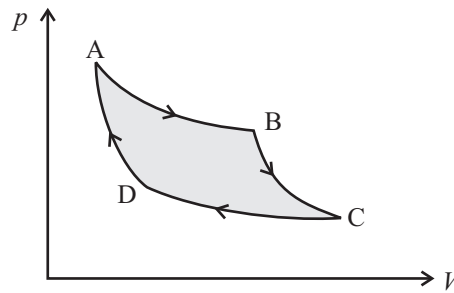


Figure 8-17: Carnot cycle in the p - V diagram for an ideal gas as the working substance; starting from the state A, the gas is made to undergo a cyclic process in four stages: an isothermal expansion at a temperature T_1 to state B, an adiabatic expansion to state C in which the temperature falls to T_2 , an isothermal compression at temperature T_2 to state D lying on the adiabatic curve passing through A, and finally an adiabatic compression back to state A; all the four processes are assumed to occur quasi-statically, and the entire cycle is reversible.

By making use of the equation of state of an ideal gas and the relation between the pressure, volume, and temperature of the gas in an adiabatic process (sec. 8.10.6.1), one can arrive at the following relation,

$$\frac{Q_2}{Q_1} = \frac{T_2}{T_1}, \quad (8-54a)$$

which immediately gives, for the efficiency of a Carnot engine,

$$\eta = 1 - \frac{T_2}{T_1}. \quad (8-54b)$$

Problem 8-10

Derive the relation (8-54a).

Answer to Problem 8-10

HINT: Let the pressure and volume of the gas in states A, B, C, D in fig. 8-17 be (p_1, V_1) , (p_2, V_2) , (p_3, V_3) , and (p_4, V_4) respectively. In an isothermal process involving a given quantity of an ideal gas, the internal energy of the gas remains constant. Hence the heat absorbed (Q_1) in the isothermal expansion at temperature T_1 has to be the same as the work performed by the gas, i.e., $Q_1 = \nu RT_1 \ln \frac{V_2}{V_1}$, where ν stands for the number of moles of the gas. Similarly the heat rejected by the gas in the isothermal compression at temperature T_2 is $Q_2 = \nu RT_2 \ln \frac{V_3}{V_4}$. Considering the four stages of the cyclic process described above, we have $p_1 V_1 = p_2 V_2$, $p_2 V_2^\gamma = p_3 V_3^\gamma$, $p_3 V_3 = p_4 V_4$, $p_4 V_4^\gamma = p_1 V_1^\gamma$ (see sec. 8.21.4, especially eq. (8-102)), from which we get $\frac{V_2}{V_1} = \frac{V_3}{V_4}$ (check this out). This leads to eq. (8-54a).

One observes from the relation (8-54b) that, the lower the temperature of the heat sink in comparison with the temperature of the heat source, the higher is the efficiency of the engine. If it were possible for the working substance to give up heat to a sink at the absolute zero of temperature, the efficiency would be unity - the dream efficiency for any engine. However, the principles of thermodynamics impose a stricture against cooling any system down to the absolute zero of temperature - a stricture that derives from the third law of thermodynamics introduced in sec. 8.14.1. Moreover, for practical reasons an actual engine seldom works with the heat sink maintained at a temperature less than the atmospheric temperature. This puts a limit on the *ideal efficiency* of the engine. Finally, the efficiency of an actual engine working on the basis of a cycle operating between two given temperatures T_1 and T_2 (i.e., one with two isothermal and two adiabatic stages in a complete cycle) will be less than the ideal efficiency given by expression (8-54b) because of the dissipative effects mentioned above.

For a Carnot cycle run with a working substance other than an ideal or a real gas, the graphical representation of the cycle will look different compared to that in fig. 8-17, where the thermodynamic state diagram may involve variables other than the pressure and volume. Regardless, the expression for the efficiency for the cycle, run between temperatures T_1 and T_2 , will always be given by eq. (8-54b). A *universal* way to represent

a Carnot cycle is in terms of the variables T and S (the temperature and the entropy) of the working substance when it looks like the rectangular figure shown in fig. 8-18.

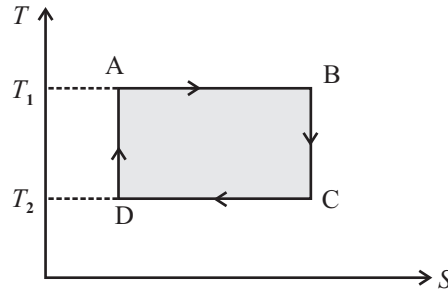


Figure 8-18: Carnot cycle in the T - S diagram; in the case of an engine run with an ideal gas as the working substance, the states A, B, C, D correspond to those in fig. 8-17.

Actual engines employed for the practical purpose of power generation are seldom based on the Carnot cycle since the extraction (as also the rejection) of heat by the working substance at a single constant temperature often poses problems of a practical nature (see, for instance, sec. 8.15.6). Instead, the working substance is made to accept heat effectively from sources with temperatures varying over a range. As a result, the ideal efficiencies of these engines (arrived at by ignoring heat dissipation in the thermodynamic cycles followed in these) get reduced compared to the efficiency of a Carnot engine working under comparable conditions.

While the construction and running of an actual engine involves a large number of practical considerations, of greater interest in the present context are the thermodynamic cycles underlying the operations of these engines where these cycles are idealized ones, based on the assumption that dissipative processes are absent, i.e., in other words, the cycles are *reversible*.

Problem 8-11

Show that a reversible engine drawing heat from and rejecting heat to a series of reservoirs with maximum and minimum temperatures, respectively, T_{\max} and T_{\min} cannot have an efficiency greater than that of a Carnot engine operating between these two temperatures.

Answer to Problem 8-11

HINT: For the sake of simplicity, we assume that the temperatures of the sources and sinks for the reversible engine (call it R) in question, make up a discrete set - the generalization to continuously ranging sets of temperatures is straightforward. Among the sources, (say, M in number) consider any one with temperature, say, τ_i ($i = 1, 2, \dots, M$) from which R draws Q_i amount of heat, and similarly, consider a representative source at a temperature τ_α ($\alpha = 1, 2, \dots, N$, say), to which R rejects Q_α amount of heat. Considering a complete cycle of operation of R involving all the sources and sinks mentioned, let W be the work delivered by it.

Imagine now a number of additional Carnot engines C_i ($i = 1, 2, \dots, M$) and C_α ($\alpha = 1, 2, \dots, N$) such that C_i draws heat q_i from the source at temperature T_{\max} and rejects Q_i to the system at τ_i (this acts as a source for R but as a sink for C_i); C_α , on the other hand, draws Q_α from the system at τ_α and rejects heat q_α to the sink at T_{\min} (fig. 8-19). According to the given condition, we have $T_{\max} \geq \tau_i, \tau_\alpha \geq T_{\min}$ for all i, α .

Imagining a complete cycle of operation of R, along with complete cycles, specified above, of all the additional Carnot engines, we observe that all the participating systems are returned to their initial states excepting the systems at T_{\max}, T_{\min} since a net amount of heat $\sum_i q_i$ is drawn from the former and, likewise, $\sum_\alpha q_\alpha$ is rejected to the latter and, at the same time, an amount of work $w = W + \sum_i (q_i - Q_i) + \sum_\alpha (Q_\alpha - q_\alpha)$ is delivered by the composite cycle, which thus is equivalent to a Carnot cycle. Thus,

$$\frac{W + \sum_i (q_i - Q_i) + \sum_\alpha (Q_\alpha - q_\alpha)}{\sum_i q_i} = 1 - \frac{T_{\min}}{T_{\max}}.$$

But, $q_i = \frac{T_{\max}}{\tau_i} Q_i$, and $q_\alpha = \frac{T_{\min}}{\tau_\alpha} Q_\alpha$, and hence,

$$W = \sum_i Q_i \left(1 - \frac{T_{\min}}{\tau_i}\right) - \sum_\alpha Q_\alpha \left(1 - \frac{T_{\min}}{\tau_\alpha}\right),$$

(check this out). Hence the efficiency of the engine R ($\frac{\text{work delivered}}{\text{total heat absorbed}}$) is

$$\eta_R \equiv \frac{W}{\sum_i Q_i} \leq \frac{\sum_i Q_i \left(1 - \frac{T_{\min}}{\tau_i}\right)}{\sum_i Q_i},$$

i.e., $\eta_R \leq 1 - \frac{T_{\min}}{T_{\max}}$. Assuming that at least one of the temperatures τ_i ($i = 1, 2, \dots, M$) is less than T_{\max} , or at least one of the temperatures τ_α is larger than T_{\min} , check that the last statement reduces to a strict inequality.

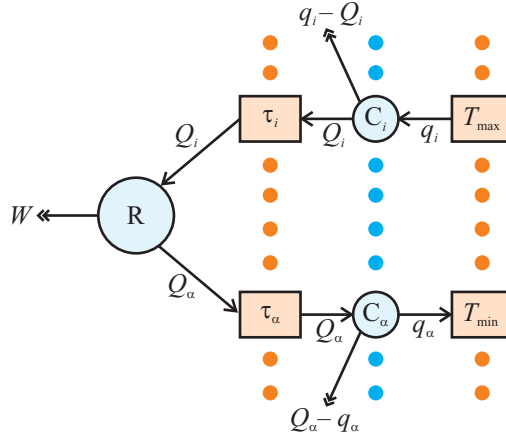


Figure 8-19: Comparing the efficiency of a reversible engine (R) drawing heat from and rejecting to reservoirs with temperatures distributed over a range, with the efficiency of a Carnot cycle working between the maximum and minimum temperatures (T_{\max}, T_{\min}) of that range; there are M number of sources with temperatures τ_i ($i = 1, 2, \dots, M; \tau_i \leq T_{\max}$) of which one is shown, from which R draws Q_i amount of heat; similarly, R rejects heat Q_α to a sink at temperature τ_α ($\alpha = 1, 2, \dots, N; \tau_\alpha \geq T_{\min}$); two auxiliary heat reservoirs at temperatures T_{\max}, T_{\min} are imagined; C_i and C_α are additional Carnot cycles ($i = 1, 2, \dots, M, \alpha = 1, 2, \dots, N$) imagined to operate such that in a combined complete cycle of all these engines all the sources at τ_i, τ_α are returned to their initial states, and the net result is that of a Carnot engine operating between T_{\max}, T_{\min} ; double-headed arrows indicate work delivered by the various engines.

At the cost of repeating myself, I summarize below the various limitations on the efficiencies of ideal and actual heat engines

1. The second law decrees that heat extracted from a body cannot be converted into work without leaving any other change in the universe. This requires the efficiency to be less than unity:

$$\eta < 1. \quad (8-55a)$$

2. Working between two fixed temperatures $T_1, T_2 (< T_1)$, an engine, drawing heat at the higher temperature has to reject a part of it at the lower temperature. The maximum possible efficiency under this condition is attained (in a limiting sense) by a reversible engine, the operation of which is described by a Carnot cycle with efficiency

$$\eta_C = 1 - \frac{T_2}{T_1} < 1. \quad (8-55b)$$

3. If an engine R draws heat from and rejects to reservoirs with temperatures distributed over a range from, say, T_2 to T_1 ($T_2 < T_1$), then its efficiency will be less than that a Carnot engine operating between T_1, T_2 , even if R operates in a reversible cycle,

$$\eta_R < \eta_C = 1 - \frac{T_2}{T_1} < 1. \quad (8-55c)$$

4. Finally the efficiency (η) of an actual engine drawing heat from and rejecting to reservoirs at specified temperatures, will always be less than that of a reversible engine (R) drawing from and rejecting to the same set of reservoirs since the actual engine always involves dissipation due to irreversible processes,

$$\eta < \eta_R. \quad (8-55d)$$

Among the various thermodynamic cycles (commonly referred to as ‘power cycles’) made use of in the practical field of power generation, we shall have a brief look at the Rankine cycle, the Otto cycle, and the Diesel cycle (sections 8.15.6 and 8.15.7) - three cycles of great interest since they are involved in the operations of a conventional power plant, a petrol engine, and a Diesel engine respectively.

8.15.4 The absolute scale of temperature

Part of what I have said in sec. 8.15.3 is based on subtle logical considerations that I have left implied. The principles of thermodynamics imply that the efficiency of a Carnot engine depends *only* on the temperatures of the heat source and the heat sink and not, for instance, on the nature and the amount of the working substance used. The efficiency, then, is a universal feature of all Carnot engines running between a source and a sink at given temperatures. This fact, along with the relation (8-54a), can be made use of in order to define an *absolute scale of temperature*, also referred to as the *thermodynamic scale* or the *Kelvin scale*. It is the absolute scale (or, what is effectively the same thing from the operational point of view, the SI scale) that has been used in referring to temperatures throughout this book (refer, for instance, to sections 8.4.3,

8.14,). We can now have a look at the theoretical basis of this temperature scale.

Strictly speaking, the relations (8-54a) and (8-54b) should hold with the temperatures expressed in the *gas scale*, an empirical temperature scale based on the properties of an ideal gas (see sec. (8.10.5) for an introduction to the gas scale), since these are derived by referring to a Carnot cycle run with an ideal gas as the working substance. Indeed, it is the temperature in the gas scale that primarily appears in the equation of state of an ideal gas (see sec. 8.10.6). In earlier sections I used the gas scale and the absolute scale interchangeably based on a presumed identity between the two scales (see, once again, sec. 8.10.5). It is now time to examine why this identity should hold.

Considering a Carnot cycle operated with an arbitrarily chosen working substance, the above result of the independence of the efficiency on the nature and amount of the working substance tells us that the relation (8-54a) must continue to apply, with the temperatures once again expressed in the gas scale. But now one can make use of the result that the value of $\frac{Q_2}{Q_1}$ depends only on the temperatures of the source and the sink to define a *new* scale of temperature, independently of the nature of the working substance, such that the ratio $\frac{Q_2}{Q_1}$ equals the ratio of the temperatures in this *new* scale. Evidently, then, the ratio of any two temperatures in the new scale agrees with the corresponding ratio in the gas scale. Thus, the two scales can be made to agree entirely with each other by making their numerical values the same for any one chosen temperature. This is done by assigning the value 273.16 to the temperature of the triple point of water in both the scales (corresponding to the choice of the constant α in sec. 8.10.5).

8.15.5 Scales of temperature: summary

The concept of temperature as a thermodynamic variable for any and every thermodynamic system is arrived at from the zeroth law of thermodynamics. While the zeroth law provides a sound logical foundation for the concept of temperature, the idea of temperature follows empirically as well, from experience and observations. The question then comes up of *measuring* the temperature of any given body in quantitative and nu-

merical terms. This can be done in various ways, leading to various different empirical temperature scales. An empirical temperature scale depends on some specific thermodynamic property of some particular thermodynamic system. While differing in their reference to these specific aspects, all empirical temperature scales are to be consistent with one another since, in the ultimate analysis, they measure in numerical terms the same physical quantity.

The gas scale is an empirical temperature scale, based on the properties of an ideal gas which can be realized in practice by making use of a real gas and employing appropriate conversion formulae to allow for the difference in the behavior of a real gas and an ideal gas.

Finally, the absolute thermodynamic scale is a universal temperature scale, not tied to the properties of any particular thermodynamic system. By its definition, and by the choice of the numerical value of the temperature of the triple point of water, it is made to coincide with the gas scale. The SI scale of temperature, or the Kelvin scale, is a practical or operational realization of the thermodynamic scale where different but specific empirical scales are made use of in various different temperature ranges, and appropriate conversion formulae are employed.

8.15.6 The Rankine cycle

A conventional power plant makes use of a mixture of water and water vapor as the working substance, where the working substance is made to *flow* from one part of the plant to another in the process of going through a complete thermodynamic cycle, an idealization of which is the *Rankine cycle*. The composition of the working substance (i.e., the proportion of water and water vapour) varies as it flows from one part of the plant to another during the cycle, the principal components of the plant being the boiler, the turbine, the condenser and the feed pump. Since the working substance receives heat in the boiler and delivers work to the prime mover at the turbine., the power plant differs essentially from the petrol and Diesel engines, the latter being *internal combustion engines* where the working substance (a gas mixture) is heated in a closed chamber and

the expansion of the gas for the delivery of work to the prime mover occurs in the *same* chamber. Considering the nature of the working substance, the Rankine cycle is referred to as a *vapor power cycle* in contrast to the Otto and Diesel cycles (pertaining respectively to the petrol and the Diesel engines), the latter being *gas power cycles*. You will find some of the concepts relating to change of state, relevant to the understanding of a Rankine cycle, discussed in section 8.22.

The basic Rankine cycle is shown schematically in a p - V diagram in fig. 8-20, where the working substance is made to pass successively through the states marked 'a', 'b', 'c', 'd', and 'e', the cycle being completed as the working substance returns to the state 'a'. In this diagram, the state 'a' corresponds to saturated liquid (i.e., liquid at the boiling point under a given pressure) at a high pressure inside the boiler, where it receives heat at a constant pressure and temperature, and is gradually converted into steam, the state 'b' corresponding to the entire mass of the working substance being in the form of saturated vapor. During this process of boiling under a high pressure, the pressure and temperature remains constant while, the volume changes along with the composition of the working substance.

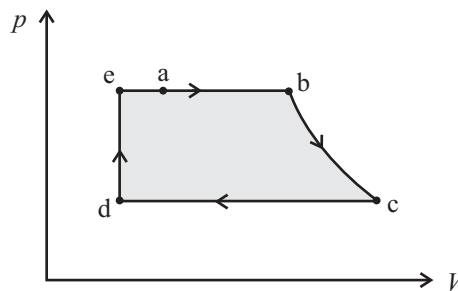


Figure 8-20: The Rankine cycle in the p - V diagram (schematic); a mass of water is boiled at a constant pressure and temperature from the state 'a' to the state 'b'; the resulting steam is made to expand adiabatically from state 'b' to state 'c' in which the working substance is a mixture of water and water vapour; the next stage is a process of condensation at a constant pressure and temperature to water in state 'd'; the water is then raised to boiler pressure (state 'e') and, finally, taken back to state 'a' by being heated to the boiler temperature; the entire cycle is assumed to be reversible.

The steam in the state 'b' is made to flow to the turbine assembly where it issues from a set of narrow nozzles and impinges on a set of blades fitted to the turbine rotor -

the prime mover on which work is performed - by the adiabatic expansion of steam. This process is represented by the part of the cycle from the state 'b' to 'c', where the latter corresponds to *wet steam* (a mixture of steam and water droplets) issuing from the turbine assembly. In practice, however, the saturated steam from the boiler is *superheated* (the process of superheating is not shown in the figure) before being sent to the turbine assembly, so that at the end of the adiabatic expansion one obtains dry rather than wet steam.

The next stage of the cycle corresponds to heat being given out by the working substance in the condenser, where the process of condensation is made to proceed up to the state 'd', the latter corresponding to saturated water at the condenser pressure. This water is then raised to the boiler pressure in the feed pump (state 'd' to 'e') and then heated in the boiler till it becomes saturated at the boiler pressure (state 'a'), i.e., its temperature rises to the corresponding boiling point.

If the same working substance were made to pass through a Carnot cycle, then the process of condensation would have to be terminated at a particular point before the state 'd' is arrived at. Such a termination of the condensation process at a preassigned point is not possible practically and, moreover, the working substance would then be a mixture of water and steam instead of saturated water. Raising this mass to the boiler pressure would then pose a huge problem in the design of the feed pump.

The stage of the cycle from 'e' to 'a' corresponds to the heating of the working substance in a process in which its temperature increases over a large range, i.e., the heating is not isothermal as in a Carnot cycle, and can be described effectively (allowing for some idealization) as a process in which the working substance absorbs heat from a succession of sources maintained at different temperatures. As a result, the ideal efficiency of a Rankine cycle works out to a value less than that of a Carnot cycle working between two temperatures equal to the maximum and the minimum temperatures attained in the Rankine cycle.

8.15.7 The Otto and Diesel cycles

The Otto and the Diesel cycles are idealized thermodynamic cycles describing the basic processes taking place in a petrol engine and a Diesel engine respectively. The working substance in either of these two engines is a mixture of air and a fuel (petrol vapor or sprayed Diesel oil) during part of a cycle and a mixture of air and the combustion products resulting from the fuel during the other part.

The heating of the working substance in either of these two cycles does not occur by a process of heat transfer between a high temperature source (a hot flame or, more generally, a heated mixture of air and combustion products generated in a heating chamber) and the working substance but rather appears directly in the latter as energy released in the chemical process of combustion. Moreover, the working substance is renewed periodically by discharging it after the completion of a cycle and drawing in a fresh supply for the next cycle. All these features make it difficult to represent the processes in the petrol or the Diesel engine faithfully in a thermodynamic state diagram. It is therefore customary to refer to *air-standard* Otto and Diesel cycles where it is assumed that a certain mass of air is made to pass through an equivalent set of processes as in the case of the actual working substance. Finally, one adopts the usual idealization of assuming that all the processes are reversible, as in the Rankine and Carnot cycles.

The petrol and Diesel engines are referred to as *reciprocating* engines since, in these, mechanical energy is imparted primarily in the form of a to-and-fro motion which is then converted to a rotary motion as necessary. Each of these is commonly operated as a *four stroke* engine, where the outward and the inward movements of the piston fitted in the cylinder containing the working substance are made up of four stages of movement or strokes in a complete cycle (two stroke engines are also possible). Two of these correspond to the intake of the gas mixture (the induction stroke) at the beginning of a cycle, and the discharge of the used up mixture (the exhaust stroke) at the end of the cycle. In the air-standard Otto and Diesel cycles, however, these two strokes are not represented, since the working substance in these two cycles is assumed to be an equivalent amount of air. This does not cause any major problem in the theoretical

analysis of the working of the engines since these two strokes are not of much relevance from the thermodynamic point of view.

8.15.7.1 The Otto cycle

Fig. 8-21 is a schematic representation of the air-standard Otto cycle in a p - V diagram. Starting from the state 'a', the working substance (air; in reality, a mixture of air and petrol vapor is drawn in in an earlier *charging stroke* from a chamber termed the *carburettor*) is compressed adiabatically to 'b', when its volume gets reduced by the *compression ratio* $r = \frac{V_1}{V_2}$ from V_1 to V_2 .

Next, the working substance gets heated from 'b' to 'c' at constant volume. In the actual petrol engine, a spark is ignited in the cylinder containing the gaseous mixture, when a combustion takes place, releasing chemical energy that heats up the gas at approximately a constant volume. The working substance (which, in an actual engine is now a mixture of air and the combustion products) then expands adiabatically ('c' to 'd') to the original volume V_1 , the expansion ratio being thus the same as the compression ratio. Finally, heat is rejected at constant volume ('d' to 'a') to complete the cycle where, in the actual engine, this involves a partial discharge of the gas mixture to the outer atmosphere. The complete discharge and the subsequent intake for the next cycle involves one more to-and-fro motion of the piston.

The idealized efficiency of the Otto cycle is given by the expressions

$$\eta = 1 - \left(\frac{1}{r}\right)^{\gamma-1} = 1 - \frac{T'}{T}, \quad (8-56)$$

where γ stands for the ratio of the two specific heats ($\gamma = \frac{c_p}{c_v}$) of the working substance (see sec. 8.21.4). The temperatures T' and T in the second expression are the temperatures of the states 'd' and 'c' respectively shown in fig. 8-21. Note that, while T is the highest temperature attained by the working substance during a cycle, T' is *not* the lowest temperature since the cooling continues up to the state 'a', when the temperature

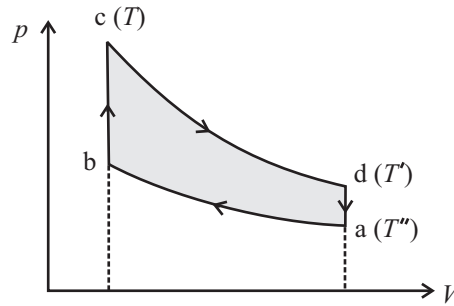


Figure 8-21: The air-standard Otto cycle; a mass of air, assumed to be equivalent to the actual working substance, is taken through a cyclic process made up of an adiabatic compression (state 'a', volume V_1 , to state 'b', volume V_2), heating at a constant volume from state 'b' to state 'c', adiabatic expansion from state 'c' to state 'd', and finally, cooling at constant volume back to state 'a'; the compression ratio is $r = \frac{V_1}{V_2}$; the entire cycle, assumed to be a reversible one, is a useful approximate representation of the processes occurring in a petrol engine.

becomes, say, T'' . Thus the efficiency of a Carnot cycle working between the highest and lowest temperatures attained in a cycle would have been $1 - \frac{T''}{T}$, which is higher than the efficiency of the Otto cycle. As already mentioned, this is because of the fact that heating and cooling in the Otto cycle are not done isothermally.

8.15.7.2 The Diesel cycle

Fig. 8-22 depicts schematically the air-standard diesel cycle in which the thermodynamic processes effectively correspond to those occurring in a Diesel engine. Once again, the charging and the exhaust strokes of the latter are not represented in the air-standard cycle.

Starting from the state 'a', the working substance (air; in an actual Diesel engine, the working substance in this part of the cycle consists of a mixture of air and sprayed Diesel oil) is compressed adiabatically up to the state 'b' from a volume V_1 to V_2 , the compression ratio (r) being $\frac{V_1}{V_2}$.

In reality, the Diesel oil is sprayed not at the beginning of the compression process, but only after a substantial degree of compression has been achieved.

The compression is followed by the addition of heat to the working substance *at a con-*

stant pressure ('b' to 'c'). In the actual Diesel engine, the mixture of air and sprayed Diesel oil reaches the ignition point because of the rise in temperature resulting from the compression. The pressure remains almost constant during the heating since the Diesel oil is added to the engine cylinder at a controlled rate. In some engines, however, there takes place a rise in the pressure during the early parts of the combustion. The ratio $\frac{V_3}{V_2}$ (see fig. 8-22) is termed the *cut-off ratio* (ρ).

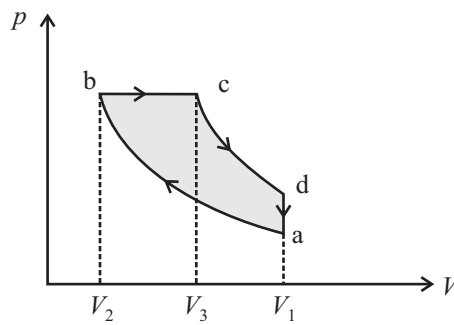


Figure 8-22: The air-standard Diesel cycle; a mass of air, assumed to be equivalent to the actual working substance, is taken through a cyclic process made up of an adiabatic compression (state 'a', volume V_1 , to state 'b', volume V_2 ; compression ratio $r = \frac{V_1}{V_2}$), heating at a constant pressure from state 'b' to state 'c' for which the volume is V_3 (cut-off ratio $\rho = \frac{V_3}{V_2}$), adiabatic expansion from state 'c' to state 'd', and finally, cooling at constant volume back to state 'a'; the entire cycle, assumed to be a reversible one, is a useful approximate representation of the processes occurring in a Diesel engine.

The heating of the working substance is followed by an adiabatic expansion up to the original volume V_1 ('c' to 'd'). Finally, heat is rejected at a constant volume till the cycle is completed. In the actual Diesel engine, this involves partial discharge of the mixture of air and the combustion products as in the case of the petrol engine.

The idealized efficiency of the Diesel cycle is given in terms of the compression- and the cut-off ratios by the expression

$$\eta = 1 - \frac{1}{\gamma r^{\gamma-1}} \frac{\rho^{\gamma} - 1}{\rho - 1}, \quad (8-57)$$

where, once again, γ stands for the ratio of the two specific heats of the working substance.

The efficiency of the Diesel cycle turns out to be less than that of the Otto cycle with the same compression ratio. When the efficiency of the Diesel cycle is compared to that of a Carnot cycle with the same maximum and minimum temperatures attained by the working substance during a cycle, the former is seen to be less than the latter, which is a consequence of the fact that the addition and the rejection of heat in the Diesel cycle do not occur isothermally.

8.15.8 Refrigeration

A refrigerator is, in a sense, a heat engine run in the reverse. Its function is to extract heat from a body at a comparatively low temperature (the body to be cooled), and further to receive energy in the form of work from an external source, giving up heat to a thermal reservoir at a higher temperature (commonly, the atmosphere). This process, run in successive cycles, achieves the desired function of lowering the temperature of the body under consideration to an appropriately low value. Fig. 8-23 depicts schematically the energy exchange process undergone by the working substance of a refrigerator.

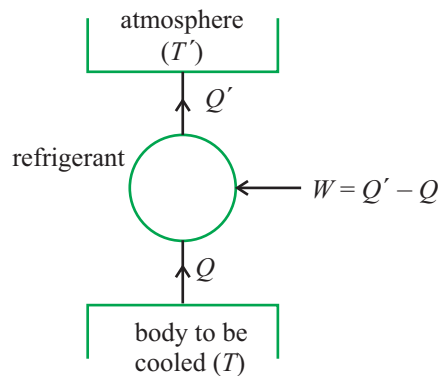


Figure 8-23: Energy flow diagram for a refrigerator; a certain amount of a refrigerant absorbs Q amount of heat from the substance to be cooled, receives W amount of energy in the form of work, and delivers Q' amount of heat to a heat reservoir (commonly, the atmosphere) at a higher temperature.

Fig. 8-24 is a block diagram of the components of a refrigeration system. The *refrigerant*, a liquid with a low boiling point and a high latent heat of evaporation (see sec. 8.22.2),

is made to pass from one component to another in this system in a cyclical manner, where it goes through alternating processes of *evaporation under a low pressure and condensation under a relatively high pressure*, thereby making possible the process of refrigeration.

The extraction of heat from the body to be cooled occurs with the refrigerant in the evaporator where it is made to evaporate, extracting the latent heat of evaporation from the body. The evaporator is a low pressure section of the refrigeration system, facilitating the evaporation process. As the refrigerant gets evaporated, it is drawn into the compressor wherein its pressure and temperature are increased to appropriately high values by the performance of work on the refrigerant. The compressor then releases the vapor raised to the high pressure and temperature into the condenser, housed outside the refrigeration chamber where the vapor gets cooled by giving up heat to the atmosphere and condenses partially to the liquid phase. The fluid then expands adiabatically into the evaporator through an expansion valve, where the refrigerant begins a new cycle.

The performance of the refrigerator is expressed quantitatively by the *work ratio*

$$r = \frac{Q}{W}, \quad (8-58)$$

where Q stands for the heat extracted in a cycle from the body to be cooled, and W for the amount of work performed per cycle on the refrigerant (see fig. 8-23). Assuming that the refrigerant passes through a Carnot cycle in the reverse direction, the work ratio is given by the expression

$$r = \frac{T}{T' - T}, \quad (8-59)$$

where T and T' are, respectively, the temperatures at which the extraction and rejection of heat take place. The designing of the refrigerator aims at achieving a comparatively high value of r .

Air conditioning is based on essentially the same principle as refrigeration.

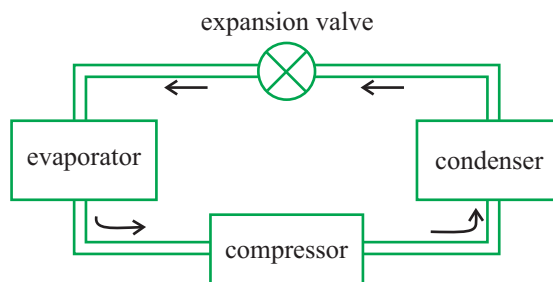


Figure 8-24: Block diagram of the various components in a refrigeration system.

The choice of an appropriate *refrigerant* is important for the efficient and relatively non-hazardous operation of a refrigerator. Among other things a good refrigerant has to have a low boiling point and a high latent heat of vaporization. Additionally, it has to be relatively less corrosive and has to meet certain requirements relating to environmental factors. Chlorofluorocarbons, which were commonly used till recent decades, are now being replaced with ammonia, sulphur dioxide, and non-halogenated hydrocarbons because of the role of the former in ozone depletion.

Problem 8-12

100 cal of heat is to be extracted from a bowl of water at 278K in a refrigerator, assumed to operate reversibly, which releases heat to the atmosphere at temperature 305K. How much of work is to be performed on the refrigerant in order to achieve this?

Answer to Problem 8-12

SOLUTION: The required work is obtained from equations (8-58) and (8-59) as $W = Q \frac{T' - T}{T}$, where $Q = 420.0\text{J}$ (refer to sec. 8.8.1; we assume $J \approx 4.2\text{J}$ per cal), $T = 278\text{K}$, and $T' = 305\text{K}$. This gives $W = 40.8\text{J}$ (aprox).

8.16 Thermal expansion of solids, liquids, and gases

Observations and experience tell us that all substances possess the property of *thermal expansion* regardless of whether they are solids, liquids or gases. If the temperature of a body is increased at constant pressure, its volume is found to increase.

In the case of a gas, the explanation of thermal expansion can be sought in the kinetic theory. With an increase in temperature the average speed of the gas molecules increases. Consequently, imagining a small planar area around any given point, the mean rate of momentum transfer through that area will increase if the number density of molecules remains constant. Conversely, for a given rate of momentum transfer, i.e., for a given pressure, the number density of molecules, and hence the density of the gas has to decrease with increasing temperature, implying a volume expansion.

The kinetic theory of liquids is somewhat more complicated as compared to gases. A qualitative and partial explanation of the thermal expansion of a liquid is based on *density fluctuations* in the liquid. The average distance between the molecules of a liquid being much less than that of a gas, the liquid molecules are clustered in some regions, and sparsely distributed in some others, with comparatively large gaps in between. The clusters break up continually due to molecular collisions and are formed anew, exchanging positions with the sparsely populated regions. With an increase in temperature, collisions become more frequent, the clusters break up more quickly, and the average volume of the clusters decreases, i.e., the average gaps between the molecules increases. This leads to an increase in volume of the liquid. The constraint of constant pressure has a less important role to play here compared to that in the case of a gas.

In the case of *water*, however, the effect of temperature on the mean spacing among molecules is somewhat different, especially at temperatures close to its freezing point. This I will address briefly in a later section (sec. 8.19.2).

In the case of a solid, the molecules are not mobile like those of a liquid or a gas. The mean spacing between the molecules of a solid is much smaller compared to that of a liquid or a gas. Consequently, a molecule interacts rather strongly with other molecules

surrounding it. As a result of such interactions among the molecules, any particular molecule, on the average, remains pinned at a certain location inside the solid without being able to move about throughout its volume like the molecules of a liquid or a gas. If it gets displaced from the pinned position, it experiences an unbalanced force drawing it back to that position, as a result of which the molecules *vibrate* around their respective mean positions. In other words, while the molecules possess *translational* motion in a fluid, they can undergo only *vibrations* in a solid.

At low temperatures, the amplitudes of the vibrations are small and the unbalanced force acting on a displaced molecule along each of the three co-ordinate axes resembles the restoring force acting on a particle in simple harmonic motion. At higher temperatures, on the other hand, the amplitudes are also higher and the force on a molecule differs from the linear restoring force responsible for simple harmonic motion. The motion now becomes *anharmonic* and the mean position of the molecule gets displaced from what it was in harmonic motion at lower temperatures. In addition, the frequency of oscillation about the mean position acquires an amplitude dependence. The condition of minimum energy (or more precisely, a thermodynamic function referred to as *free energy*) at any given temperature then relates the mean separation between the atoms with the temperature, implying a thermal expansion.

Another distinctive feature of thermal expansion of solids as compared to liquids and gases relates to the fact that a solid has a definite shape as well as a volume. This means that one can speak of the expansion along a particular direction or in any planar section of the solid, as also of an expansion in volume. A liquid or a gas, on the other hand, does not have any definite shape, as a result of which the concepts of linear and surfacial expansion are of no relevance.

The linear, surfacial, and volume expansions of a solid are not all independent of one another, though. The definitions of the three *coefficients of expansion* will be given below and the relation between these will also be worked out in the case of an isotropic solid.

The thermal expansion of a liquid differs from that of a gas in that the magnitude of

expansion for a given increase in temperature is much less for a liquid as compared to a gas. Indeed, the volume expansion of the liquid is usually comparable to that of its containing vessel, as a result of which one has to distinguish between the *apparent* and *real* expansions of the liquid. The magnitude of expansion of a gas, on the other hand, being much larger, the volume expansion of the containing vessel is entirely negligible and hence one need not, in practice, distinguish between apparent and real expansions.

8.17 Thermal expansion of solids

8.17.1 Coefficients of expansion of a solid

As a solid undergoes thermal expansion, it expands equally in all directions. Consequently, knowing the expansion along any chosen direction, one can work out the expansion in any plane section of the solid as also the expansion in volume.

Imagine any linear stretch of length l in the solid at a temperature T . Suppose that, as a result of expansion, the length of this stretch changes to $l + \delta l$ as the temperature is made to change to $T + \delta T$. Then the proportional increase in length for an increase δT in the temperature is

$$\text{proportional expansion} = \frac{\text{increase in length}}{\text{initial length}} = \frac{\delta l}{l}. \quad (8-60)$$

The proportional increase in length per unit rise in temperature is then

$$\alpha = \frac{\delta l}{l \delta T}. \quad (8-61)$$

Observations tell us that this quantity does not depend appreciably on the temperature T (at least, over a considerable range), and is a constant for a given material. It is referred to as the (thermal) *coefficient of linear expansion* of the solid under consideration.

Consider now a planar section of a body made up of the solid material under consideration. Suppose that the area of the section at a temperature T is A and that the area changes to $A + \delta A$ at temperature $T + \delta T$. Then the proportional increase in area per

unit increase in temperature works out to

$$\beta = \frac{\delta A}{A\delta T}. \quad (8-62)$$

This, once again, is a constant for the solid material under consideration, and is referred to as the *coefficient of surfacial expansion* of the solid.

Finally, suppose that the volume of a body made up of the solid material under consideration is V at a temperature T , and that the volume changes to $V + \delta V$ at temperature $T + \delta T$. The proportional increase in volume per unit increase in temperature is then

$$\gamma = \frac{\delta V}{V\delta T}. \quad (8-63)$$

A constant for the material under consideration, γ is termed its *coefficient of volume expansion*.

Since the coefficients of linear, surfacial, and volume expansion of solids do not vary appreciably with temperature, it is customary to define these by referring to the length, surface area, and volume respectively, at 273K.

8.17.2 Relation between the three coefficients for a solid

Fig. 8-25 depicts a planar section S of a body made of an isotropic solid, the area of the section being A at temperature T , as in eq. (8-62). This area can be thought of as being made up of a large number of small squares, a number of such squares being shown in the figure. The squares cover the entire area A of the section, except for irregular shaped regions (like, say, the region marked R in the figure) along the border. The total area of these regions is, however, negligible compared to A . Let the edge length of a typical square at temperature T be a , which changes to, say, $a + \delta a$ at temperature $T + \delta T$. Then, according to the definition of the coefficient of linear expansion, one has

$$\alpha = \frac{\delta a}{a\delta T}. \quad (8-64)$$

Again, the area of the elementary square under consideration is $A = a^2$ at temperature T , which changes to $A + \delta A = (a + \delta a)^2$ at temperature $T + \delta T$. For a sufficiently small value of δT , the square and higher powers of δa can be ignored (as compared to a^2), and thus the proportional change in area of the square can be approximated as $\frac{\delta A}{A} = \frac{2\delta a}{a} = 2\alpha\delta T$ (check this out using (8-64)).

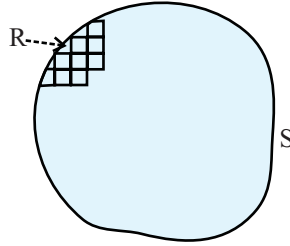


Figure 8-25: A planar section S through a body made of a solid material; the section can be partitioned into a large number of squares of infinitesimal size, a number of which are shown in the figure; the squares cover the entire area of the section except for the small irregular shaped regions along the border, one such region being R ; the total area of these irregular regions is negligible compared to A ; as the solid expands with a rise in temperature from, say, T to $T + \delta T$, the edge length of a typical elementary square increases from a to $a + \delta a$, consequent to which the area of the square also increases.

Note that this proportional change in area of a square depends only on the change in temperature δT and the value of the constant α , and not on the edge length a of the square. Hence, the proportional increase in area of all the squares is the same, being also the proportional increase in area of the entire section S . Moreover, the proportional increase in area is seen to be the same for all sections through the solid body under consideration. The proportional increase in area per unit rise in temperature, i.e., the coefficient of surfacial expansion thus works out to (see eq. (8-62))

$$\beta = \frac{1}{\delta T} 2\alpha\delta T = 2\alpha. \quad (8-65)$$

Consider now the body referred to above, of volume, say, V at temperature T , made of the solid material. Regardless of the shape of the body, it can be partitioned into a large number of elementary cubes, each of vanishingly small volume. The cubes cover the volume V except for small irregular shaped regions along the boundary surface of

the volume. For a vanishingly small volume of the elementary cubes, however, the total volume of these irregular regions turns out to be negligibly small compared to V . For a typical cube of edge length a at temperature T , let the edge length be $a + \delta a$ at temperature $T + \delta T$. The change in volumes of the elementary cube is then $(a + \delta a)^3 - a^3 \approx 3a^2\delta a$, where, once again, the squares and higher powers of δa have been neglected, assuming δT to be sufficiently small. The proportional increase in volume of the elementary cube is then $\frac{3\delta a}{a} = 3\alpha\delta T$, where the definition (8-61) has been made use of. This being the same for all the elementary cubes, is also the proportional increase in volume ($\frac{\delta V}{V}$) of the whole body. The coefficient of volume expansion is then given by

$$\gamma = \frac{\delta V}{V\delta T} = 3\alpha. \quad (8-66)$$

In summary, we arrive at the following relation between the three coefficients of thermal expansion of an isotropic solid

$$\alpha = \frac{\beta}{2} = \frac{\gamma}{3}. \quad (8-67)$$

Problem 8-13

A rod of length $l = 0.50$ m expands in length by $\delta l_1 = 1.0 \times 10^{-3}$ m when heated through a certain temperature difference. Another rod, made of a different material, and of the same length l expands by $\delta l_2 = 6.0 \times 10^{-4}$ m when heated through the same temperature range. Parts of the two rods are sawed off and joined end to end to make up a composite rod of length l once again, when it is found that the composite rod expands in length by $\delta l = 7.0 \times 10^{-4}$ m when heated, once again, through the same temperature range. What are the lengths taken from the two rods to make up the composite rod?

Answer to Problem 8-13

HINT: The coefficients of linear expansion of the two materials are $\alpha_1 = \frac{\delta l_1}{l\delta T}$ and $\alpha_2 = \frac{\delta l_2}{l\delta T}$, where δT stands for the temperature difference referred to in the statement of the problem. If the lengths taken from the two rods be xl and $(1-x)l$ so as to make up the length l of the composite rod, then

$\delta l = \alpha_1 x l \delta T + \alpha_2 (1 - x) l \delta T = x \delta l_1 + (1 - x) \delta l_2$, where the above expressions for α_1 and α_2 are made use of. This gives $x = \frac{\delta l - \delta l_2}{\delta l_1 - \delta l_2} = \frac{1}{4}$. In other words, the lengths taken from the two rods are 0.125 m and 0.375 m.

Notice that the value of δT is not necessary for working out the required lengths ($x l$ and $(1 - x) l$). All that is needed here is that the lengths of the two rods and of the composite rod are to be the same and that, moreover, δT is to be the same for all the three.

8.17.3 Instances of thermal expansion of solids

1. Railway tracks made of steel get heated during the daytime and cooled during the night. Moreover, the average temperature of the tracks increases in the summer and decreases in the winter. This fluctuation in temperature causes a variation in the length of the tracks. In order to eliminate the possibility of bending and buckling of the tracks due to thermal expansion, the tracks are divided into segments, with small gaps left at more or less regular intervals between segments. As the tracks lengthen with a rise in the temperature, the gaps close up, opening again as the temperature falls.
2. The working principle of a *bi-metallic strip* is based on the thermal expansion of solids. A bimetallic strip is made up of two metallic strips A and B of equal length but made of dissimilar metals, firmly riveted on to each other (fig. 8-26(A)). Supposing that the coefficient of linear expansion of the material of the strip A is larger than that of B, what does one expect to happen as the temperature of the strip is made to increase?

If the two metallic strips were not firmly attached to each other, then A would have expanded more than B and the combination would look as in fig. 8-26(B). But since the two are firmly attached, the bimetallic strip will get bent as in fig. 8-26(C). In the present instance, the strip B will be the inner member while A will lie outside because of the larger expansion of the latter. If, on the other hand, the temperature is decreased, the bimetallic strip will bend the other way.

This bending of a bi-metallic strip can be made use of in setting up a *temperature sensor*. An electrical circuit is made up with the bimetallic strip included in such a way that the circuit gets disconnected if the strip gets bent due to a change in the temperature away from a desired level. As the strip straightens back, the circuit gets connected again.

3. Circular metallic strips are firmly mounted on wooden wheels of bullock carts and carriages so as to make these more durable. Before a strip, in the form of a ring, is mounted on a wheel, it is heated so as to expand in diameter, allowing it to be easily mounted. On getting cooled, the ring contracts and tightly presses on the wheel, guarding it and giving it firmness against impacts.

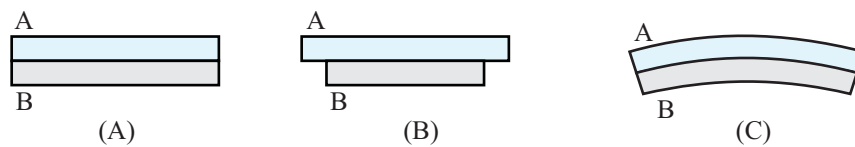


Figure 8-26: A bi-metallic strip; (A) strips A and B of equal lengths but made of dissimilar metals are riveted on to each other; (B) if the strips were not attached firmly, the two strips would change in length independently as the temperature is made to increase; (C) due to the riveting, the combination will bend on heating; a lowering of temperature will produce a bend in the opposite direction.

8.18 Thermal stress

if a body is constrained in such a way that its thermal expansion or contraction is resisted by some means, then a *stress* is set up in it such that the associated strain tends to cancel the expansion or contraction that would have taken place in the absence of the constraint. This is what is referred to as *thermal stress*.

For instance, if a metallic rod is firmly clamped at both ends on to rigid supports, then a longitudinal compressive stress is set up in it as its temperature is made to increase. On the other hand, an elongational stress is developed on a lowering of the temperature. Elongational and compressive stresses are distinguished from one another with a positive or a negative sign respectively in the value of stress.

Suppose that the length of the rod is l at a temperature T , and that the temperature is changed by δT . If the rod were not constrained, its change in length would have been $\delta l = \alpha l \delta T$ (see eq. (8-61)). However, the rod being clamped, an equal and opposite elastic change in length has to take place. This means that the elastic strain has to be $-\frac{\delta l}{l} = -\alpha \delta T$. Consequently, the stress developed will be given by

$$\begin{aligned} \text{longitudinal stress} &= \text{Young's modulus of the material} \times \text{longitudinal strain} \\ &= -Y\alpha\delta T. \end{aligned} \tag{8-68}$$

The negative sign in this expression indicates that the thermal stress has to be a compressive one for an increase in the temperature.

Evidently, the bending of a railway track with a rise in the temperature is a consequence of thermal stress being developed in the tracks, similar to that in a rod clamped at both ends. In the case of the rod, if the increase in temperature be sufficiently large, the resulting compressive stress causes the rod to *buckle*, which is what happens to the railway track if no gaps are left between successive segments.

The bending of a bi-metallic strip due to a change in temperature, or the tightening of a heated ring on the rim of a cart wheel as the ring gets cooled are also instances of the development of thermal stress.

If a cylinder containing a gas is heated to a sufficient temperature, it bursts. Here the cylinder may be assumed to be of a constant volume since the walls of the cylinder do not expand appreciably. The gas inside the cylinder, on being constrained to a constant volume, develops a high pressure due to the increase of temperature. This is, in reality, an instance of thermal stress since the pressure in a fluid is nothing but the bulk stress, taken with a negative sign.

At times, the thermal stress in a body may be due to a *non-uniform* thermal expansion. For instances, if hot water is poured in a glass jar with a thick wall, the jar may crack. The inner layer of the glass wall becomes suddenly heated in contact with the hot water

while the outer layer remains colder because of the fact that heat is conducted only slowly through the thick wall. Hence, a stress develops in the wall of the jar due to unequal expansions of the two layers. As the stress exceeds a tolerable limit, the jar cracks.

Problem 8-14

A body is heated from $T = 273\text{K}$ to $T + \delta T = 283\text{K}$, and simultaneously a pressure is applied on it from all sides such that its volume remains unchanged. If the bulk modulus of the material of the body is $K = 1.6 \times 10^{11}$ Pa, and its coefficient of volume expansion is $\gamma = 4.0 \times 10^{-5}\text{K}^{-1}$, what is the pressure necessary?

Answer to Problem 8-14

SOLUTION: If no pressure were applied, the volume expansion δV would be $\delta V = \gamma V \delta T = 4.0 \times 10^{-5} \times V \times 10\text{m}^3$. The pressure required must be such that this expansion is neutralized. Since an equal pressure (say, p) applied from all sides is equivalent to a compressive volume stress p , a compression of volume by δV , i.e., a compressive volume strain of $\frac{\delta V}{V}$ would imply $\frac{p}{\frac{\delta V}{V}} = K$, where K is the bulk modulus. Thus $p = \frac{\delta V}{V} \times 1.6 \times 10^{11}$ Pa, i.e., $4.0 \times 10^{-4} \times 1.6 \times 10^{11}$ Pa.

8.19 Thermal expansion of liquids

8.19.1 Apparent expansion and real expansion of a liquid

In measuring the volume expansion of a liquid due to a rise in temperature, one has to keep in mind the fact that there takes place, at the same time, an increase in volume of the vessel in which the liquid is contained. As a result, one needs to distinguish between the *apparent* expansion and the *real* expansion of the liquid.

For the sake of concreteness, suppose that the liquid under consideration is contained in a cylindrical vessel whose area of cross section is A . Let, at a temperature T , the height of the liquid surface above the bottom of the vessel be h . In other words, the volume of the liquid at temperature T is $V = Ah$. For the sake of convenience, let us

assume that a mark has been put against the liquid surface on the wall of the vessel (see fig. 8-27).

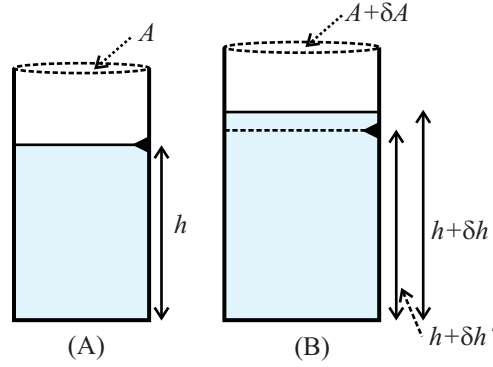


Figure 8-27: Illustrating the apparent and real expansions of a liquid; the containing vessel is assumed to be of a cylindrical shape; (A) the liquid level at temperature T , against which a mark has been put on the wall of the vessel, at a height h above the bottom; the area of cross-section of the vessel is A ; (B) the liquid level at temperature $T + \delta T$; the position of the mark has changed due to the expansion of the vessel; at the same time, the area of cross-section of the vessel has also changed; the apparent expansion is obtained by disregarding the change in height of the mark and also the change in cross-section.

Let now the temperature be changed by a small amount to $T + \delta T$, as a result of which the height of the liquid surface above the bottom of the vessel changes to $h + \delta h$. Due to the change in temperature, there has to occur a change in the area of cross-section of the cylindrical vessel. This change would have been the same if the cylinder were solid instead of a hollow one, and is given by the expression (see equations (8-62) and (8-67))

$$\delta A = A\beta_s\delta T = \frac{2}{3}A\gamma_s\delta T, \quad (8-69)$$

where β_s and γ_s denote the coefficients of surfacial and volume expansion respectively, of the material of the cylindrical vessel.

The volume of the liquid at the temperature $T + \delta T$ being $V + \delta V = (A + \delta A)(h + \delta h)$, the real change in volume due to the change δT in temperature is

$$\delta V = A\delta h + h\delta A, \quad (8-70)$$

where the product of the two small quantities δA and δh has been ignored.

Strictly speaking, one has to consider the limit $\delta T \rightarrow 0$ in all the above expressions, starting from the definitions of the coefficients of thermal expansion, in which case the end results arrived at by ignoring small terms become exact ones. In actual measurements, however, δT and the respective expansions are small but finite, and there arises the possibility of an error when the theoretical results are compared with observed ones. These errors, however, are entirely negligible in most situations of interest.

Let us now see what the *apparent* expansion of the liquid works out to. Suppose that the mark placed on the wall of the vessel at the height h at temperature T , has reached up to a height $h + \delta h'$ at the altered temperature $T + \delta T$. This change in height occurs due to the linear expansion of the material of the vessel. One thus has (see equations (8-61) and (8-67))

$$\delta h' = h\alpha_s\delta T = \frac{1}{3}h\gamma_s\delta T, \quad (8-71)$$

where α_s is the coefficient of linear expansion of the material of the vessel.

The apparent rise in the height of the liquid column in the vessel is $\delta h - \delta h'$. At the same time, in calculating the apparent expansion of the liquid, i.e., the expansion of the liquid without taking into account the expansion of the vessel, an observer will disregard the change in the cross-section of the vessel. The apparent change in volume is thus

$$\delta V' = A(\delta h - \delta h'). \quad (8-72)$$

The coefficients of real and apparent expansion of the liquid are defined as

$$\gamma_r = \frac{1}{V} \frac{\delta V}{\delta T}, \quad (8-73a)$$

and

$$\gamma_a = \frac{1}{V} \frac{\delta V'}{\delta T}, \quad (8-73b)$$

respectively. The relation between the coefficients of real and apparent expansion then works out to

$$\gamma_r - \gamma_a = \gamma_s, \quad (8-74)$$

(check this out).

8.19.2 The anomalous expansion of water

The thermal expansion of water is of a somewhat exceptional character as compared to the expansion of most other liquids. It has been observed that in the temperature range from 0° C to 4° C water undergoes volume *contraction* rather than expansion on heating. Above 4° C, however, the expansion of water is similar in nature as that of other liquids. The volume of a given mass of water is found to be minimum at 4° C, corresponding to which its density is maximum.

In looking for an explanation for this anomalous behavior of water, one notes that the molecules of water are weakly tied to each other by a certain type of bonding known as the *hydrogen bond*. Because of this bonding, the molecules tend to form clusters, with sparsely populated gaps in between the clusters. In contrast to cluster formation in other liquids, the formation of these clusters requires that the molecules should have a chance, so to speak, to flock together during their thermal motion, which in turn requires that their mobility should not be too low. Close to 0° C, the mobility happens to be so low that clusters are relatively rare. The formation of clusters by hydrogen bonding increases as the temperature is increased toward 4° C, above which, however, the *break-up* of clusters due to thermal motion starts to predominate and the thermal expansion resembles that in other liquids.

As the temperature is made to decrease below 0° C, the molecules completely lose their

mobility and water freezes into ice. One observes a considerable similarity between the structure of ice crystal and the disposition of water molecules close to 0°C . In both the structures, there occurs small clusters of molecules with comparatively larger gaps in between. It may be mentioned in this connection that water is characterized by another anomalous property, namely, it *expands on freezing*. In other words, the density of ice is *less than* that of water. This is closely related to the anomalous thermal expansion of water between 0°C and 4°C . As a sample of ice melts, water molecules, owing to their mobility, partly fill up the gaps in the crystalline structure of ice, and thus there occurs a decrease of the volume.

Effect on marine life in freezing weather

As the water of a lake cools down at sub-zero temperatures it is the uppermost layer of water that gets cooled first, when its density increases and it moves down, displacing the relatively lighter layers near the bottom of the lake. The latter, on coming to the surface, gets cooled, and as its temperature drops sufficiently it, in turn, moves down. This process of relative displacement of various parts of a fluid due to a difference in temperature is termed *convection* (see section 8.23.2). As the temperature of the atmosphere drops, the water in the lake gets cooled, with the bottom layers remaining at a lower temperature compared to the upper layers. However, as the temperature of the layer near the bottom reaches 4°C it can no longer be displaced by the water in the upper layers even as the latter gets cooled below 4°C , because the density *decreases* below 4°C . Thus, the temperature of the bottom layer continues to remain at 4°C while the upper layers get cooled and finally the lake starts freezing from the top. Ice being a bad conductor of heat, thermal conduction cannot take place from the relatively warm bottom layer out to the atmosphere. As a result, even when the atmospheric temperature drops considerably below the freezing point and the upper layer of the lake remains frozen, the bottom layer at a temperature of 4°C keeps on supporting marine life.

8.20 Thermal expansion of gases

While talking of thermal expansion of a solid or a liquid, we do not usually make a separate mention of the pressure. The expansion usually takes place under the atmospheric pressure, and this pressure remains unchanged in the change of temperature and volume. In other words, the coefficients of thermal expansion defined for solids and liquids are commonly the ones corresponding to *the pressure being held constant*.

In the case of thermal expansion of gases, on the other hand, the pressure has to be taken into account as a relevant variable. It is often the case for a gas that there occurs a change in volume *and* pressure as a result of a change in the temperature. For the sake of convenience, I will assume that the gas under consideration obeys the ideal gas equation.

The coefficient of thermal expansion of the gas *at constant pressure* is defined as

$$\gamma_p = \frac{1}{V} \left[\frac{\delta V}{\delta T} \right]_{p = \text{constant}}, \quad (8-75)$$

where δV stands for the change in the volume (V) for a change δT in the temperature, and where the constancy of pressure has been indicated because, if the pressure were allowed to change, the change in volume would have been different.

If now one makes use of the ideal gas equation ($pV = \nu RT$), one finds that for a constant pressure p ,

$$\left[\frac{\delta V}{\delta T} \right]_{p = \text{constant}} = \frac{\nu R}{p}. \quad (8-76)$$

One then finds that the coefficient of thermal expansion at constant pressure is the reciprocal of the temperature T :

$$\gamma_p = \frac{1}{T}. \quad (8-77)$$

The expression $\left[\frac{\delta V}{V} \right]_{p = \text{constant}}$ stands for the proportional change in volume at constant

pressure. If the change in volume is expressed in relation to the volume (say, V_0) at 0° C, then one gets, instead of (8-77),

$$\gamma_p^{(0)} = \frac{1}{V_0} \left[\frac{\delta V}{\delta T} \right]_{p = \text{constant}} = \frac{1}{T_0}, \quad (8-78)$$

where T_0 is the absolute temperature corresponding to 0° C, i.e., $t_0 = 273K$ (approx).

The coefficient of thermal expansion at constant pressure ($\gamma_p^{(0)}$) defined in this manner is thus seen to be a constant:

$$\gamma_p^{(0)} = \frac{1}{273} \text{ K}^{-1} \text{ (approx)}. \quad (8-79)$$

In reality the gas under consideration may not exactly follow the ideal gas equation, which may make $\gamma_p^{(0)}$ depend on the temperature to a small extent.

In this connection, mention may be made of the *pressure coefficient of a gas at constant volume*. It is defined as the rate at which the proportional change in pressure at a constant volume varies with the temperature:

$$\gamma_V = \frac{1}{p} \left[\frac{\delta p}{\delta T} \right]_{V = \text{constant}}, \quad (8-80)$$

where, for the sake of clarity, the constancy of volume has been indicated. On making use of the equation of state for an ideal gas, one finds that γ_V is given by the same expression as γ_p :

$$\gamma_V = \frac{1}{T}. \quad (8-81)$$

In defining the pressure coefficient at constant volume, one often expresses the proportional change in pressure with reference to the pressure at 0° C, i.e., replaces $\frac{\delta p}{p}$ with $\frac{\delta p}{p_0}$. One then obtains, in place of eq. (8-81),

$$\gamma_V^{(0)} = \frac{1}{p_0} \left[\frac{\delta p}{\delta T} \right]_{V = \text{constant}} = \frac{1}{T_0}, \quad (8-82)$$

which implies

$$\gamma_V^{(0)} = \gamma_p^{(0)} = \frac{1}{273} \text{ K}^{-1} \text{ (approx).} \quad (8-83)$$

The definitions of γ_p and γ_V and the distinction, in principle, between the two, are relevant for solids and liquids as well. As I have already mentioned, the coefficients of thermal expansion of solids and liquids as commonly defined, refer implicitly to a constant pressure, i.e., correspond to γ_p defined above. One could also define γ_V for a solid or a liquid as the proportional increase in pressure, required to keep the volume constant, per unit increase in the temperature. Indeed, because of the constraint of constant volume, this relates to the thermal stress introduced in section 8.18. The value of this coefficient for a solid or a liquid is usually much larger than that for a gas.

8.21 Calorimetry

The theory and methods underlying the measurement of the energy exchanged in the form of heat by a system with other systems is referred to as *calorimetry*. As the system under consideration receives heat, its temperature increases, while the temperature decreases as it rejects heat to other systems (exceptions to this general rule will be discussed in sec. 8.22). The quantity of heat received or given away by a body can thus be determined by measuring the rise or drop in the temperature of the body.

8.21.1 Thermal capacities of bodies

Supposing that the change in the temperature of a body is δT as it absorbs δQ amount of heat in a quasi-static process, the ratio $\frac{\delta Q}{\delta T}$ is referred to as the *thermal capacity* (also referred to as the *heat capacity*) of the body. However, one should at the same time mention the physical condition under which the exchange of heat takes place. For instance, if the pressure on the body remains constant during the exchange, then $\frac{\delta Q}{\delta T}$, i.e., the heat absorbed by the body per unit rise in temperature, is termed the thermal

capacity *at constant pressure*:

$$C_p = \left[\frac{\delta Q}{\delta T} \right]_{p = \text{constant}}. \quad (8-84)$$

In the case of a solid or a liquid, p usually stands for the atmospheric pressure, which remains constant as the body exchanges heat. Hence the term heat capacity for a solid or a liquid usually refers to C_p defined above, and the constancy of pressure is sometimes left implied.

On the other hand, if the *volume* of the body under consideration remains constant during the process of heat exchange, then the ratio $\frac{\delta Q}{\delta T}$ is referred to as the thermal capacity *at constant volume*:

$$C_V = \left[\frac{\delta Q}{\delta T} \right]_{V = \text{constant}}. \quad (8-85)$$

As I have mentioned above, the heat capacity at constant pressure (C_p) is usually the quantity of relevance for a solid or a liquid, where the symbol C without a suffix is sometimes used. For a gas, on the other hand, *both* C_p and C_V are relevant, depending on circumstances. The unit of heat capacity is $\text{J}\cdot\text{K}^{-1}$, or $\text{cal}\cdot\text{K}^{-1}$.

1. In reading an equation like (8-85), one has to keep in mind that δT is a *change* in the value of the *thermodynamic variable* T , i.e., $\delta T = T_f - T_i$, where T_f and T_i are the temperatures of the final and initial states of the system under consideration. By contrast, δQ does *not* represent the change in the value of a thermodynamic quantity pertaining to the initial and final states, since it depends on the *process* under consideration which, in the case of eq. (8-84), for instance, is one at constant pressure. A more precise notation is to use the symbol $\tilde{\delta}$ to denote a small quantity representing an inexact differential. However, we omit the tilde for the sake of brevity, leaving it implied.
2. By definition, we assume δQ to be positive if heat is *absorbed* by the system, and negative if heat is *given out* by it. A negative value for the heat absorbed then implies a positive amount of heat rejected.
3. Strictly speaking, in defining the coefficients of thermal expansion, one has to

consider the limit $\delta T \rightarrow 0$, in which case the magnitudes of thermal expansions are also vanishingly small. In actual observations and measurements, however, this limit is not realized, though it can be approximated closely.

4. It is a common practice in thermal physics to designate *extensive* variables like volume and internal energy with upper case symbols like V and U . *Intensive* variables (see sec. 8.9) like the pressure (p), on the other hand, are designated with lower case symbols. An exception to this practice is made for the temperature (T) in the absolute scale. The heat capacities C_p and C_V are extensive quantities, while the *specific* heat capacities (specific heats, in brief) c_p and c_V defined below are intensive quantities.

8.21.2 Specific heats of substances

The heat capacities of bodies made of the same substance varies in proportion to their masses, i.e., are extensive quantities. The heat capacity *per unit mass* is referred to as the *specific heat* of the material the body is made of. Denoting the heat capacity by the symbol C (with an appropriate suffix as required), the specific heat $\frac{C}{m}$, is denoted by the symbol c (again with an appropriate suffix as required) or, at times by s . Evidently, the unit of specific heat is $\text{J}\cdot\text{kg}^{-1}\cdot\text{K}^{-1}$ or, in mixed units, $\text{cal}\cdot\text{kg}^{-1}\cdot\text{K}^{-1}$.

At times, one defines the specific heat by referring to the heat capacity per *mole*, rather than per unit *mass*. In other words if the heat absorbed for an increase in temperature of ν mol of a substance by δT be δQ , then the specific heat is given by

$$\text{(at constant pressure)} \quad c_p = \frac{1}{\nu} \left[\frac{\delta Q}{\delta T} \right]_{p = \text{constant}}, \quad (8-86a)$$

$$\text{(at constant volume)} \quad c_V = \frac{1}{\nu} \left[\frac{\delta Q}{\delta T} \right]_{V = \text{constant}}. \quad (8-86b)$$

These are sometimes referred to as *molar* specific heats. The unit of molar specific heat is $\text{J}\cdot\text{mol}^{-1}\cdot\text{K}^{-1}$ or, in mixed units, $\text{cal}\cdot\text{mol}^{-1}\cdot\text{K}^{-1}$.

For ordinary or even moderately low temperatures, say, not lower than 100 K, the specific

heats of most substances are found to be independent of the temperature. In that case, if a body of mass m or mole number ν be heated from a temperature, say, T_1 to T_2 , then the heat required will be given by

$$Q = mc(T_2 - T_1), \quad (8-87a)$$

where c stands for the heat capacity per unit mass (with appropriate suffix depending on the context), or

$$Q = \nu c(T_2 - T_1), \quad (8-87b)$$

where c stands for the heat capacity per mole (again, with appropriate suffix depending on the context).

8.21.3 Specific heats of an ideal gas

In the last section (sec. 8.21.2), I have defined the specific heats at constant pressure and at constant volume (c_p and c_V respectively) by equations (8-86a) and (8-86b). While talking of the specific heat, I will for the time being refer to the molar specific heat for the sake of concreteness.

There is nothing special about the constraints of constant pressure and constant volume, excepting that these correspond to commonly occurring experimental situations.

One can, in principle, consider any *other* constraint such as, say, a constant value of p^2V if you like, and consider the ratio $\frac{\delta Q}{\delta T}$ under *this* condition. This will then define a 'specific heat', though it may not have much relevance in practice.

Following these definitions of c_p and c_V , one can work out a useful relation between the two specific heats of an *ideal gas*.

For a quasi-static process connecting two states of a fluid close to each other in the

thermodynamic state space, the first law of thermodynamics reads

$$\delta Q = \delta U + p\delta V, \quad (8-88)$$

which is simply eq. (8-10), with δQ written in place of $\tilde{\delta}Q$ for the sake of brevity. If now the volume of the gas remains unchanged during the process connecting the two states, one has $[\delta Q]_{V=\text{constant}} = \delta U$, and thus,

$$c_V = \frac{\delta u}{\delta T}, \quad (8-89)$$

where I have used the *molar internal energy* u in place of U (considering just one mole of the gas as our system of interest), which is an intensive thermodynamic variable. If the fluid under consideration be an *ideal gas* then the condition $V = \text{constant}$ need not be mentioned separately on the right hand side of eq. (8-89) because the internal energy of any given amount of an ideal gas (1 mole in the present instance) does not actually depend on the volume. Using now eq. (8-28) for the internal energy of the ideal monatomic gas and setting $\nu = 1$ (which gives $u = \frac{3}{2}RT$), one gets

$$c_V = \frac{3}{2}R. \quad (8-90)$$

The experimentally measured values of molar specific heats of monatomic gases at moderate or high temperatures are found to agree quite well with this ideal gas value.

Similar considerations apply for the specific heat of an ideal gas at constant pressure. The equation of state for one mole of an ideal gas being $pV = RT$ (for the sake of consistency, one should write v for V here because one is referring to the molar volume, which is an intensive quantity, but this distinction is commonly overlooked) one has, for any two states lying close to each other in the state space,

$$p\delta V + V\delta p = R\delta T, \quad (8-91)$$

where the symbol δ denotes the difference in values of the relevant quantity for the two states under consideration. If, now the pressure remains constant during a process

then, putting $\delta p = 0$, one gets

$$p\delta V = R\delta T. \quad (8-92)$$

The equation (8-88) expressing the first law of thermodynamics connecting the two states then takes the form

$$\delta Q = \delta u + R\delta T. \quad (8-93)$$

Since we are considering one mole of the gas, we ought to have used the symbol δq in place of δQ , but once again, I prefer not to have to use too many symbols.

The ratio $\frac{\delta Q}{\delta T}$ under the constraint of constant pressure then works out to

$$c_p = \frac{\delta u}{\delta T} + R. \quad (8-94)$$

Once again, one need not refer to the constraint of constant pressure in the expression on the right hand side of eq. (8-94) since the internal energy of a given quantity of an ideal gas depends only on the temperature.

Finally, then, using equations (8-89) and (8-94), one arrives at the following relation between the molar specific heats c_p and c_V of an ideal gas,

$$c_p = c_V + R. \quad (8-95)$$

For a *monatomic* gas, the value (eq. (8-90)) for c_V gives

$$c_p = \frac{5}{2}R. \quad (8-96)$$

For a diatomic or a polyatomic gas, the expressions for c_V and c_p get modified due to contributions to the internal energy arising from rotational and vibrational motions of

the molecules, but the relation (8-95) remains unchanged.

The relation (8-95) can be interpreted as follows.

If the temperature of an ideal gas is increased at a constant volume, then the gas does not perform any work, and all the energy supplied to the gas in the form of heat goes to increase the kinetic energy of the molecules of the gas. On the other hand, if the gas is heated at a constant pressure, then its volume gets changed and it performs some work. This additional energy will have to be supplied to the gas in the form of heat, apart from what is required to increase the kinetic energy of the gas molecules. The increase in the kinetic energy of the molecules being the same in the two cases, c_p exceeds c_V by the amount of work performed by one mole of the gas for a unit rise in temperature at constant pressure, which our above derivation shows to be R .

The specific heat at constant pressure exceeds that at constant volume not only for a gas, but for a solid or a liquid as well. However, in the case of a solid or a liquid, the change in volume on heating at a constant pressure is usually much smaller than the corresponding change for a gas, and hence the difference in the values of the two specific heats is usually much smaller. As I have already mentioned, it is the specific heat at constant *pressure* that is commonly of relevance for a solid or a liquid.

The relations (8-90), (8-95), and (8-96) pertaining to the two specific heats of an ideal gas, are applicable for a *real* monatomic gas at relatively *high* temperatures. At lower temperatures, the behavior of a real gas departs from that implied by the ideal gas equation. At very low temperatures, i.e., close to the absolute zero, the specific heats of *all* substances become vanishingly small, and the difference $c_p - c_V$ also tends to zero.

8.21.4 Adiabatic and isothermal expansion of gases

If a gas exchanges heat with its surroundings, with its temperature kept fixed, its volume gets changed, with a corresponding change in pressure. This is an *isothermal process*. On the other hand, the volume and pressure may be changed in an *adiabatic* process in which no heat is exchanged between the gas and its surrounding systems. While an

expansion or contraction in the volume of a gas can be made to take place in various other ways (such as, for instance, in an *isobaric* process where the pressure of the gas is kept constant), isothermal and adiabatic processes are of considerable relevance from a theoretical and practical point of view.

The isothermal and adiabatic expansion of gases (the term ‘expansion’ is used for the sake of brevity to denote a volume change which, in reality, may be a contraction as well) we will consider here will be assumed to be *quasi-static* processes. I have already pointed out that an isothermal process has to be necessarily a quasi-static one. An adiabatic process, on the other, need not be necessarily quasi-static. In the present section, we will consider only quasi-static isothermal and adiabatic processes involving a gas, i.e., ones that can be represented by paths in the thermodynamic state space. The gas, moreover, will be assumed to obey the ideal gas equation of state.

For ν mol of the gas, the ideal gas equation of state reads $pV = \nu RT$. Since an isothermal process corresponds to $T = \text{constant}$, one concludes that the pressure and volume of the gas are related to each other as

$$pV = \text{constant (Boyle's law)}, \quad (8-97)$$

at every stage during an isothermal process. In other words, the variation of pressure and volume of a given quantity of a gas in an isothermal process, when represented graphically in a p - V diagram, will correspond to a rectangular hyperbola. Fig. 8-28(A) shows two such curves, corresponding to isothermal processes at two different temperatures, say, T_1 and $T_2(> T_1)$.

During a quasi-static adiabatic process, on the other hand, the pressure and volume follow a different relation. For such a process, where no heat exchange takes place between the gas and its surroundings, the first law of thermodynamics, as expressed by eq. (8-10), gives, with $\delta Q = 0$ (recall that we have agreed to omit the tilde in $\tilde{\delta}Q$),

$$\delta U + p\delta V = 0. \quad (8-98)$$

One can now make use of the equation (8-89) to write, for ν mol of the gas, $\delta U = \nu \delta u = \nu c_V \delta T$ to obtain

$$p\delta V + \nu c_V \delta T = 0. \quad (8-99)$$

At the same time, the ideal gas equation, on differentiation of both sides, gives

$$p\delta V + V\delta p = \nu R\delta T, \quad (8-100)$$

where the symbols δp , δV , etc., all denote infinitesimally small quantities. The last two equations taken together, along with eq. (8-95), imply, for an adiabatic process,

$$\frac{c_p}{c_V} \frac{\delta V}{V} + \frac{\delta p}{p} = 0, \quad (8-101)$$

(check this out).

The ratio of the two specific heats ($\frac{c_p}{c_V}$) of a gas is commonly denoted by the symbol γ . On performing integration on both sides of the last equation, one obtains the relation between the pressure and volume during a quasi-static adiabatic expansion of a gas,

$$pV^\gamma = \text{constant}. \quad (8-102)$$

For a monatomic gas, equations (8-90) and (8-96) imply $\gamma = \frac{5}{3}$, while, for a diatomic gas, one has $\gamma \approx \frac{7}{5}$. Indeed, for all gases, $\gamma > 1$.

Fig. 8-28 (B) depicts a pair of adiabatic curves for a gas in the p - V diagram for two different initial states, where these curves differ in nature from the isothermal curves shown in (A). A comparison of the isothermal and adiabatic curves is shown in (C), where an isothermal curve AQB and an adiabatic curve CQD passing through any given point Q in the p - V diagram are shown. Because of the fact that the ratio of the two specific heats of the gas satisfies $\gamma > 1$, the adiabatic curve is steeper than the isothermal one.

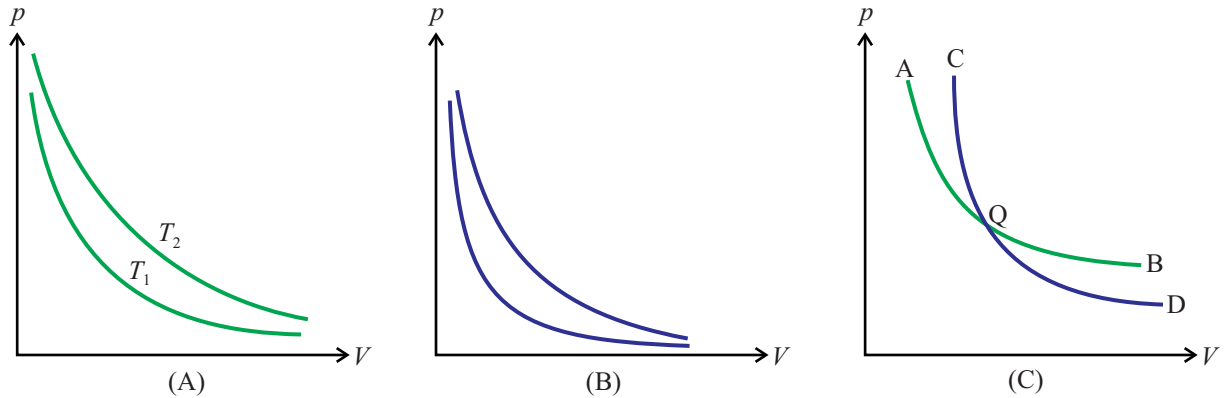


Figure 8-28: Curves (schematic) in the p - V diagram depicting isothermal and adiabatic expansions of a gas ; (A) isothermal processes at temperatures T_1 and $T_2 (> T_1)$ represented by rectangular hyperbolas; (B) quasi-static adiabatic processes with two different initial states; (C) an isothermal curve (AQB) and an adiabatic curve (CQD) passing through a given point Q in the p - V diagram; the adiabatic curve is steeper due to the fact that $\gamma > 1$.

8.21.5 The fundamental principle of calorimetry

In a calorimetric experiment, a body or a system of bodies (say, A) is made to exchange heat with another body or a system (say, B). The experiment is typically conducted in such a manner that the two systems A and B do not exchange energy in the form of work and, moreover, they do not exchange energy or matter with any other system. In reality, these conditions may be violated to a small extent, but sufficiently accurate results are often arrived at by assuming that the conditions are met with. The principle of conservation of energy (of which the first law of thermodynamics is an expression) then takes the form

$$\text{heat absorbed by system A} = \text{heat given out by system B.} \quad (8-103)$$

In reality, it may be the system A that gives out heat while B absorbs heat. As I have mentioned above, the heat *absorbed* by a system (say, Q) and the heat *given out* (say, Q') by it are, by definition, equal and opposite to each other ($Q' = -Q$). This means that eq. (8-103) is valid in general.

In this formula, the quantity of heat absorbed or given out can be worked out from equations (8-87a) and (8-87b). Together, these formulae are said to constitute the

fundamental principle of calorimetry.

1. In the case of a system made up of a number of bodies, the formula for heat absorbed is a straightforward generalisation of equation (8-87a) or (8-87b). For instance, for a system made of two bodies of masses m , m' and specific heat c , c' (say, per unit mass) respectively, eq. (8-87a) is generalized to

$$Q = mc(T_2 - T_1) + m'c'(T'_2 - T'_1), \quad (8-104)$$

where T_1 , T'_1 are the initial temperatures of the two systems and T_2 , T'_2 are the respective final temperatures.

2. The formulae for heat absorbed and heat rejected are to be modified if any of the substances involved in the heat exchange undergoes a *change of state*, where the *latent heat* of that substance comes in. This I am going to explain sec. 8.22.

Usually, a calorimetric experiment is performed with a view to determining the specific heat or the latent heat (see section 8.22) of a substance.

Problem 8-15

Of two bodies A and B, the former gets cooled at the rate of $r_1 = 0.5 \text{ K}\cdot\text{s}^{-1}$, and the heat given out by A goes to raise the temperature of B. If the masses of the two bodies be $m_1 = 0.05 \text{ kg}$ and $m_2 = 0.04 \text{ kg}$ respectively, and the specific heats of the respective materials be $s_1 = 400 \text{ J}\cdot\text{kg}^{-1}\cdot\text{K}^{-1}$ and $s_2 = 300 \text{ J}\cdot\text{kg}^{-1}\cdot\text{K}^{-1}$, find the rate r_2 at which the body B gets heated.

Answer to Problem 8-15

HINT: The rate at which heat is given out by A and is received by B is $H = m_1 s_1 r_1$. Since this must also be given by $H = m_2 s_2 r_2$, one has, $r_2 = \frac{m_1 s_1 r_1}{m_2 s_2}$. Substituting given values, $r_2 = 0.833 \text{ K}\cdot\text{s}^{-1}$.

Problem 8-16

Imagine 1 mol of an ideal monatomic gas to be taken quasi-statically through the cyclic process depicted in fig. 8-29. Starting from the state A at pressure p_1 and volume V_1 , the gas is heated

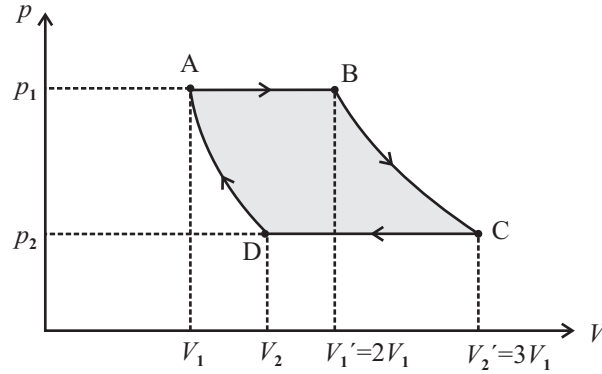


Figure 8-29: Depicting a quasi-static cyclic process undergone by 1 mol of an ideal gas; starting from the state A at pressure p_1 and volume V_1 , the gas is heated at a constant pressure to state B with volume $2V_1$, and then to undergo an adiabatic expansion to state C with volume $3V_1$; it is then cooled at a constant pressure p_2 to the state D and then taken back to A, where D lies on the adiabatic curve through A; the entire process can be looked upon as a cycle of a heat engine that absorbs heat at a varying temperature from A to B and rejects heat, once again at a varying temperature, from C to D; the highest and lowest temperatures attained during the cycle are those corresponding to states B and D respectively; since the gas does not accept and reject heat at constant temperatures, the efficiency (η) of the engine will be less than that of a Carnot engine working between the above highest and lowest temperatures (refer to problem 8-16).

at constant pressure so that the volume changes to $V_1' = 2V_1$ at state B. The gas is then made to expand adiabatically to state C with volume $V_2' = 3V_1$, after which it is cooled at constant pressure to state D that lies on the adiabatic curve through A. In the final step, the gas is compressed adiabatically so as to return to state A. Looking at the entire process as a complete cycle of a heat engine, work out the efficiency (η) of the engine. Referring to the highest and lowest temperatures attained by the gas during the cycle, and the efficiency (η') of a Carnot engine working between these two temperatures, show that $\eta < \eta'$.

Answer to Problem 8-16

HINT: Let the temperatures of the gas in states A, B, C, D be respectively T_1, T_2, T_3, T_4 . One then has, $T_1 = \frac{p_1 V_1}{R}$, $T_2 = \frac{2p_1 V_1}{R}$, $T_3 = \frac{3p_2 V_1}{R}$, $T_4 = \frac{p_2 V_2}{R}$. Considering the adiabatic process from state B to C, $p_1 (2V_1)^\gamma = p_2 (3V_1)^\gamma$, i.e., $p_2 = (\frac{2}{3})^\gamma p_1$, where $\gamma = \frac{c_p}{c_v} = \frac{5}{3}$. Let the heat absorbed by the gas in the expansion from state A to state B be Q_1 and the heat rejected in the compression from C to D be Q_2 . Then $Q_1 = c_p(T_2 - T_1) = \frac{5}{2}R(T_2 - T_1) = \frac{5}{2}p_1 V_1$, and $Q_2 = \frac{5}{2}R(T_3 - T_4) = \frac{5}{2}p_2(3V_1 - V_2)$. Considering the adiabatic process from D to A, one has, $p_1 V_1^\gamma = p_2 V_2^\gamma = (\frac{2}{3})^\gamma p_1 V_2^\gamma$, i.e., $V_2 = \frac{3}{2}V_1$. Thus, $Q_2 = \frac{15}{4}(\frac{2}{3})^\gamma p_1 V_1$. In other words, the efficiency is $\eta = 1 - \frac{Q_2}{Q_1} = 1 - \frac{3}{2}(\frac{2}{3})^\gamma = 1 - (\frac{2}{3})^{\frac{5}{3}}$.

This is less than the efficiency (η') of a Carnot engine working between the temperatures T_2 and

T_4 , the highest and lowest temperatures attained during the cycle, since $\eta' = 1 - \frac{T_4}{T_2} = 1 - \frac{p_2 V_2}{2p_1 V_1} = 1 - \frac{1}{2} \left(\frac{2}{3}\right)^{\frac{2}{3}}$.

8.22 Change of state: phase transition

If heat is added to a solid at any given pressure (p) so as to increase its temperature (T), then it will be found that at a certain temperature (say, T_M), the solid starts *melting*. If the supply of heat is continued, it will be found that the temperature remains constant at T_M while more and more of the solid melts into the liquid phase, with the entire solid being at the end converted into liquid form.

8.22.1 Change of state as phase transition

This process is an example of *phase transition*, and is referred to as melting. Here the result of adding heat to a substance does *not* appear as an increase in temperature, but is rather a change of state of the solid into the liquid at a *constant temperature*. Likewise, if heat is *extracted* from a liquid at any given pressure, then it gets cooled till it starts *freezing* into the solid form at some fixed temperature, with the entire liquid freezing over to the solid if the process of heat extraction is continued.

Other instances of phase transition involving a change of the state of aggregation of a material are *boiling* of a liquid into a gas, *condensation* of a gas into a liquid, *sublimation* of a solid into a gas, and *condensation* (or, more precisely, reverse-sublimation) of a gas into a solid. All these processes take place at a constant temperature as a result of addition or extraction of heat, if the pressure on the material is held constant at a given value. In all these processes, there takes place a noticeable change in several of the macroscopic properties, expressed in terms of the thermodynamic variables of the substance like, for instance, its density and elastic moduli.

The term *phase transition* carries a definite connotation in physics. While we shall here be concerned principally with the *changes of state* of solids, liquids, and gases, these constitute only a few instances of phase transitions of materials, there being innumer-

able *other* instances of phase transition where the state of aggregation of a material does not change.

For instance, ice is the solid form of water. In ice the molecules are arranged in a definite crystalline structure. However, this arrangement can be made to undergo notable changes by varying the pressure and temperature. Indeed, it has been found that there exist a number of different crystalline structures of ice, each differing from another in properties like the specific heat and elastic compressibility. The process of conversion from one such structure to another constitutes an instance of phase transition.

As another instance, one may consider the change in the electrical conductivity of a conducting material that takes place with a lowering of its temperature where it is seen that at a sufficiently low temperature, it becomes a *superconductor*. In this remarkable new state of the conducting material, its electrical resistance vanishes completely.

Processes of phase transition can be classified into two broad groups, namely, phase transitions of the *first* kind and those of the *second* kind, while other variants are also known. The conversion of the state of aggregation from any one among a solid, liquid or a gas to another state happens to be a phase transition of the first kind. A phase transition of the second kind is found in the transition from a liquid to a gas near the *critical point* of a material. It has been found experimentally that, at a certain temperature and pressure (taken together, these define the critical point), the commonly observed distinction between a liquid and its vapor form (like, for instance, the density difference) disappears completely. Close to the critical point, the transition from a liquid to its gaseous form involves a noticeable change in certain other finely-tuned macroscopic properties like the *density fluctuations* that show up in light scattering characteristics of the two phases.

The transition from the *paramagnetic* to the *ferromagnetic* state at the critical temperature (a specific temperature characterizing the transition) observed in some magnetic materials is another instance of a phase transition of the second kind.

I must mention here that in the transition from any one among the solid, liquid, or the

gaseous phase into another phase, or in any phase transition in general, there does *not* occur any chemical change, i.e., a change in the molecular structure of the material under consideration. What does happen is only a notable change in a number of the macroscopic (or *thermodynamic*) characteristics of the material. With reference to a change of the state of aggregation (e.g., solid, liquid or gas) of a substance, the following features are to be noted:

1. For any given value of the pressure, the transition takes place at a specific temperature, termed the *transition temperature* at that pressure. If the pressure and temperature are maintained at their respective values without supplying heat to the system or extracting heat from it, the two phases (say, solid and liquid during the process of melting) coexist with each other. On the other hand, if heat is supplied or extracted, the coexistence is disturbed and there occurs a transformation from one phase to the other at a constant temperature.
2. As the transformation takes place, there occurs a notable change in properties like the compressibility and specific volume of the substance, where the latter term refers to the volume per unit mass, i.e., the reciprocal of the density (at times, the volume per mole of the substance is termed the specific volume).
3. In order to effect the transformation of unit amount (in terms of mass or mole number) of the substance from one state to another at the given pressure, a certain definite quantity of heat is to be supplied to or extracted from it. This is referred to as the *latent heat* associated with the transformation.

8.22.2 Transition temperature. Latent heat

The transition temperature pertaining to the change of state from the solid to the liquid or from the liquid to the solid form at a given pressure is a characteristic of the material under consideration and is referred to as its melting point or freezing point respectively, these two being usually identical, especially, for a crystalline substance. The latent heat associated with such a transformation is known as the latent heat of fusion or the latent heat of solidification, depending on the process, the two being, once again, identical. Similarly, one speaks of the boiling point or the condensation point in the case

of a change of state from the liquid to the vapor or from the vapor to the liquid form. The corresponding latent heat is termed the latent heat of evaporation or condensation.

Under normal pressure (760 mm of Hg), the melting point of ice is 273 K, and its latent heat is $8.0 \times 10^4 \text{ cal}\cdot\text{kg}^{-1}$ (i.e., $3.34 \times 10^5 \text{ J}\cdot\text{kg}^{-1}$). This means that if an amount of ice is kept in contact with water at 273 K under normal pressure and the system is isolated from its surroundings, then the two will coexist in equilibrium as long as no heat is added to or taken away from it. On the other hand, if energy in the form of heat is added to the system with the pressure unchanged, then ice will get converted into water with the temperature remaining constant, the energy required for each kg of ice to melt being $3.34 \times 10^5 \text{ J}$. Conversely, this amount of energy will have to be taken out of the system for each kg of water to freeze into ice.

The boiling point of water (and condensation point of steam) under normal atmospheric pressure is 373 K and the associated latent heat is $5.37 \times 10^5 \text{ cal}\cdot\text{kg}^{-1}$ (i.e., $2.24 \times 10^6 \text{ J}\cdot\text{kg}^{-1}$). Suppose that a closed vessel contains water and water vapor at 373 K and that the pressure in the vessel is the normal atmospheric pressure. The water vapor under these conditions is said to be a *saturated* one. If the system is sealed off from the rest of the world, it will be found that the two phases in the vessel coexist with each other in equilibrium, with the amount of each phase continuing to remain unchanged.

If now an amount of heat is supplied to the system from outside, it will be found that some amount of water in the vessel gets converted to saturated vapor, which appears as bubbles throughout the volume of the water and merges with the vapor above the water surface of the vessel. If the pressure in the vessel is maintained constant by an appropriate arrangement, the temperature of the system will remain constant in the process. For each kg of water to get converted into saturated vapor in this manner, an amount of energy equal to $2.24 \times 10^6 \text{ J}$ will have to be supplied to the system.

Problem 8-17

Steam of mass $m_0 = 6.0 \times 10^{-3} \text{ kg}$ (latent heat of condensation $L_0 = 2.24 \times 10^6 \text{ J}\cdot\text{kg}^{-1}$) is made

to enter into a chamber at a temperature $T_0 = 373$ K, in which are kept $m_1 = 12.0 \times 10^{-3}$ kg of ice (latent heat of melting $L_1 = 3.34 \times 10^5$ J·kg⁻¹) and $m_2 = 6.0 \times 10^{-3}$ kg of water at $T_2 = 273$ K (specific heat 4.18×10^3 J·K⁻¹·kg⁻¹). What will be the final temperature and composition of the system? Assume that there takes place no heat exchange with any other system.

Answer to Problem 8-17

HINT: In order to solve this problem, one first needs to calculate the amount of heat given out by the steam, assuming that the entire mass m_0 condenses into water at 373 K. From the given data, this works out to $Q_0 = 1.34 \times 10^4$ J. One also needs the amount of heat necessary to melt the entire quantity of ice at 273 K, which is seen to be $Q_1 = 4.01 \times 10^3$ J. Since $Q_1 < Q_0$, the entire quantity of ice will indeed get converted into water at 273 K. Finally, one needs to know the amount of heat necessary to raise the temperature of $m_1 + m_2 = 0.018$ kg of water from 273K to 373K, which works out to be $Q_2 = 7.5 \times 10^3$ J. Since $Q_1 + Q_2 = 1.15 \times 10^4$ J, i.e., $Q_1 + Q_2 < Q_0$, the entire amount of water, after the melting of ice, will be raised to 373 K without the entire mass of steam condensing. The mass of steam condensing into water will be $m = \frac{Q_1+Q_2}{L_0} = 5.1 \times 10^{-3}$ kg. Thus the final temperature of the system will be 373 K, with 0.9×10^{-3} kg of steam remaining in the vapor state, and with $m_1 + m_2 + m = 1.8 \times 10^{-2}$ kg of water at the same temperature.

NOTE: Strictly speaking, though, this does not constitute a complete solution to the problem. The processes taking place in the chamber can be grouped into one of relatively fast heat transfer between steam on the one hand and ice and water on the other, and a slower one of evaporation or (depending on circumstances) condensation. Thus, water and steam at 373K cannot coexist unless the partial pressure of water vapor in the chamber equals the SVP of water (see sections 8.22.4, 8.22.5) at that temperature (i.e., a pressure of 1 atmosphere). In other words, there will take place either a slow evaporation of water or a slow condensation of water vapor till the condition of coexistence of water and water vapor in the chamber is met with. A system of water and water vapor in a chamber in which the air is either unsaturated or supersaturated with water vapor is only approximately an equilibrium state since there goes on slow processes referred to above till true equilibrium is reached. When such slow processes of evaporation and condensation are considered, one needs to know the volume occupied by steam in the chamber to arrive at the final equilibrium state.

Problem 8-18

The heat of combustion of a fuel is $45 \times 10^6 \text{ J}\cdot\text{kg}^{-1}$. How much of the fuel is to be burnt so as to heat 1 kg of water (specific heat $4.18 \times 10^3 \text{ J}\cdot\text{K}^{-1}\cdot\text{kg}^{-1}$) from 293K to 373K and to boil away 20% of the water (latent heat $2.24 \times 10^6 \text{ J}\cdot\text{kg}^{-1}$), if 40% of the heat released in burning is delivered to the water?

Answer to Problem 8-18

SOLUTION: If the amount of fuel required is m kg, then the heat delivered to the water is $m \times 45 \times 10^6 \times 0.40 \text{ J}$. This must be equal to $Q_1 + Q_2$, where $Q_1 = 1.0 \times 4.18 \times 10^3 \times 80$ (heat required to raise the temperature through 80 K), and $Q_2 = 1.0 \times 0.20 \times 2.24 \times 10^6 \text{ J}$ (heat required for boiling). This gives $m = 0.043$.

8.22.3 Dependence of transition temperature on the pressure

The temperature of transition in a change of state depends on the pressure, which means that the transition temperature $T(p)$ is a function of the pressure p . In the case of transition from the solid to the liquid form (or conversely) $T(p)$ usually *increases* with p . A notable exception to this rule, however, is found in the *ice-water* transition: the melting point of ice (or the freezing point of water) *decreases* with an increase of pressure. Close to the normal atmospheric pressure, the melting point of ice decreases by 0.074 K for each MPa increase in the pressure.

When a given quantity of ice melts into water, its volume *decreases*, i.e., the specific volume of ice is *larger* than that of water. For most other substances, however, the specific volume of the solid is less than that of the liquid and, at the same time, the melting points of these substances increase with pressure. In contrast, the larger value of the specific volume of ice compared to that of water is, in a way, responsible for the inverse relationship between its melting point and the pressure. The correlation between the specific volumes of the two phases involved in a transition on the one hand, and the nature of dependence of the transition temperature on the pressure on the other, can be expressed in the form of a mathematical formula, known as

Clapeyron's formula that can be derived from the laws of thermodynamics. It applies equally well to the solid-liquid, the liquid-gas, and the solid-gas transitions.

The reason why the specific volume of ice is larger than that of water, lies in the crystalline structure of ice. As I have already mentioned (see section 8.19.2), the water molecules in the ice crystal are arranged in groups, with large gaps in between. Since the molecules are not mobile in the solid phase, the gaps remain unchanged. But, as the ice melts, the molecules become mobile and tend to flock together in larger groups for short intervals of time, thereby, on the average, closing up the gaps to some extent. Consequently the average distance between the molecules in water at 0°C is somewhat less than that for ice at the same temperature, which explains why the specific volume of ice is larger than that of water.

In the case of liquid-vapor or vapor-liquid transition, however, the transition temperature is always found to increase with the pressure, there being no known exception to this rule.

The specific volume of vapor is always found to be much larger than that of the corresponding liquid at the same temperature and pressure. On making use of this fact in Clapeyron's formula, the increase of the transition temperature with pressure can be seen to follow as a consequence.

Close to the atmospheric pressure, the boiling point of water increases by roughly 1°C for a rise in pressure by 27 mm of Hg. The condensation temperature of water vapor also increases at the same rate with pressure.

8.22.4 Saturated vapor

Water and water vapor coexist with each other in equilibrium under normal atmospheric pressure at 373K. The water vapor under this condition provides an instance of a *saturated vapor*. Suppose that the boiling point of a liquid under a pressure p is $T_B(p)$, i.e.,

the liquid coexists with its vapor under a pressure p and at a temperature $T_B(p)$. Then the vapor under that pressure and at that temperature will be termed a saturated vapor.

Instead of looking at the temperature of a saturated vapor as a function of the pressure, one can express the pressure of the saturated vapor as a function of the temperature. Suppose that the pressure of a saturated vapor at a temperature T is $p_{\text{sat}}(T)$. Then $p_{\text{sat}}(T)$ is termed the *saturation vapor pressure* (SVP; also termed the saturation pressure) of the liquid at the temperature T .

8.22.5 The coexistence curve: Triple point

The dependence of the boiling point of a liquid on the pressure or, equivalently, the dependence of the saturation vapor pressure (SVP) on the temperature can be depicted graphically by a curve in a p - T diagram, with the temperature T and pressure p making up the two axes of a co-ordinate system, as in fig. 8-30. Looking at any point on the curve, the pressure (p) and temperature (T) corresponding to that point together specify a state wherein the liquid and the vapor coexist in equilibrium, the vapor under this condition being a saturated vapor. Starting from the point under consideration, if the temperature is made to increase at a constant pressure, the coexistence of the phases is disturbed and liquid gets converted into the gas while an increase of pressure at a constant temperature causes the vapor to get converted into the liquid form. Since the boiling point of a liquid rises with an increase in pressure, the slope of the liquid-vapor coexistence curve is always positive.

Such a curve is referred to as the liquid-vapor *coexistence curve* of the material under consideration. Evidently, according to the notation introduced above, one has, for any point (T, p) chosen on the curve, $p = p_{\text{sat}}(T)$, and $T = T_B(p)$.

In a similar manner, one can talk of a solid-liquid coexistence curve of a substance as in fig. 8-31. The pressure (p) and temperature (T) corresponding to any point on the curve specify a state wherein the solid and the liquid coexist in equilibrium. In general, the slope of such a solid-liquid coexistence curve is positive like that of a liquid-vapor curve,

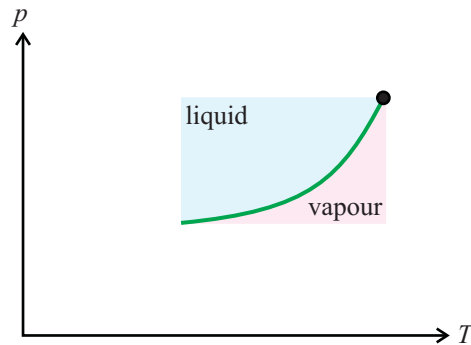


Figure 8-30: Liquid-vapor coexistence curve (schematic) in a p - T diagram; considering a typical point on this curve, the liquid and the vapor can coexist at the temperature and pressure corresponding to this point; the critical point (see sec. 8.22.1) is shown.

though the former is steeper in comparison with the latter since the melting point of a solid increases only slowly with an increase in pressure.

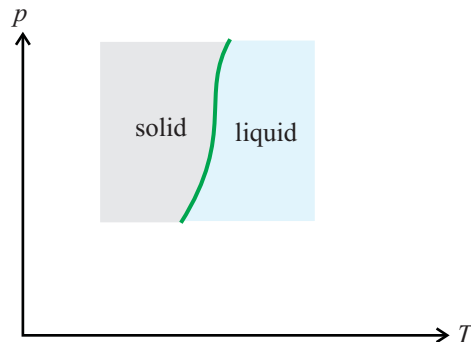


Figure 8-31: Solid-liquid coexistence curve (schematic); the curve looks similar to the liquid-vapor coexistence curve, but is steeper.

The *ice-water* coexistence curve is an exception in the sense that its slope is *negative*, since the melting point of ice decreases with an increase in the pressure (fig. 8-32).

One can, in this context, also talk of a solid-vapor coexistence curve, where a substance in the solid form co-exists with its vapor form, usually at low temperatures and pressures. Finally, one can depict all the three coexistence curves for a substance in the *same* p - T diagram, known as the *phase diagram*, as in fig. 8-33. One finds in this diagram a single point P where all the three coexistence curves meet. This means that *all*

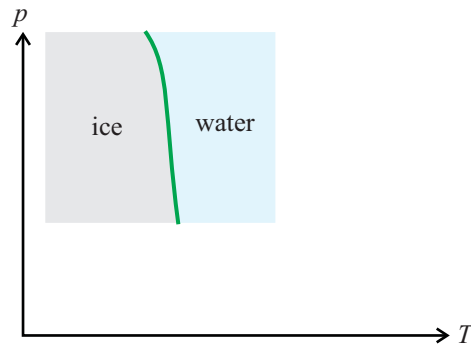


Figure 8-32: Ice-water coexistence curve (schematic), which differs from the solid-liquid coexistence curve (fig. 8-31) for most substances in that it has a negative slope.

the three phases of the substance coexist at equilibrium at the pressure and the temperature corresponding to the point P, which is thereby referred to as the *triple point* of the substance.

While the phase diagrams of most substances look qualitatively as in fig. 8-33, the phase diagram of water looks a bit different where, as mentioned above, the solid-liquid coexistence curve has a negative slope (dotted line in figure). The pressure and temperature corresponding to the triple point are $p_W = 0.61$ kPa, and $T_W = 273.16$ K. It is with reference to T_W that the SI scale of temperature is fixed.

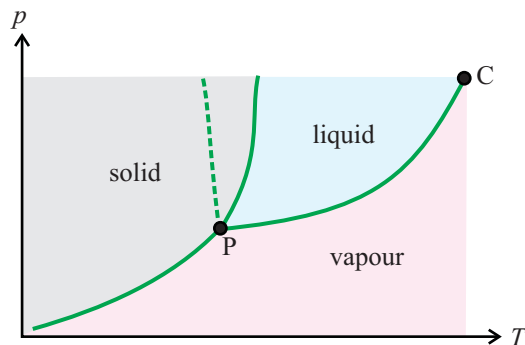


Figure 8-33: Solid-liquid-gas phase diagram; the phase diagram for a typical substance is shown, for which the specific volume of the solid form is smaller than that of the liquid form and the slope of the solid-liquid coexistence curve is positive; the triple point (P) and the critical point (C) are shown; the liquid-vapor coexistence curve does not extend beyond C; for the ice-water-steam phase diagram, the slope of the ice-water co-existence curve is negative, as depicted by the dotted line.

Notice that the liquid-vapor coexistence curve in fig. 8-33 gets terminated at the point C, termed the *critical point* for the substance under consideration, i.e., the curve does not extend beyond C, while the other two coexistence curves do not have a similar terminal point (all three curves share the triple point P as a terminal point). The critical point corresponds to a state where the distinction between a liquid and the vapor gets erased. The pressure, temperature and specific volume of a substance at the critical point are termed respectively its *critical pressure* (p_C), *critical temperature* (T_C), and *critical specific volume* (v_C). Among these three important characteristics of a substance collectively referred to as its *critical constants*, only two are independent since the three are related by the *equation of state* of the substance.

The critical constants of water are $p_C = 22.06\text{MPa}$, $T_C = 647.1\text{K}$, and $v_C = 3.125 \times 10^{-3}\text{m}^3\text{kg}^{-1}$ respectively (approximate values).

8.22.6 Gas and vapor

I have so long made no distinction between the terms ‘gas’ and ‘vapor’, using these more or less synonymously while, in reality, it is sometimes necessary to make a distinction for the sake of clarity. The term *gas* is commonly used for states at temperatures greater than T_C , while for temperatures less than T_C , the term *vapor* is preferred. As I have mentioned above, the term *saturated vapor* applies to states lying on the coexistence curve.

8.22.7 Saturated air

Suppose that the air in a room at 300 K is sealed off from the rest of the world, the pressure in the room in this condition being, say, 1.10×10^5 Pa. Suppose further that the partial pressure of the water vapor present in the room is 2.5 kPa. This is sometimes referred to simply as the *vapor pressure* of the water vapor. It is known experimentally that the saturation vapor pressure of water at 300 K is 4.5 kPa. This means that the partial pressure of water vapor in the room is less than the saturation vapor pressure at the room temperature. The air in the room is then said to be *unsaturated* (at times the term *unsaturated* is applied to the water vapor itself).

If now some water vapor is introduced into the room from outside, keeping the temperature constant, then its partial pressure will increase (the total pressure of the air in the room will increase at the same time) and at some point it will reach the saturation vapor pressure (i.e., 4.5 kPa) at the given temperature. The air in the room will then qualify as being *saturated* with respect to water vapor. If now some more water vapor at 300 K is introduced into the room, then this amount of vapor will condense into water and the remaining vapor will continue to be a saturated one. In other words, once the air becomes saturated with water vapor, it cannot hold any further water in the vapor state.

I have mentioned the value 300 K for the temperature here just for the sake of illustration. Whatever the room temperature, if the vapor pressure be less than the saturation vapor pressure at that temperature, the air in the room will qualify as unsaturated. On the other hand, if the vapor pressure becomes equal the SVP, the air will become saturated.

If, instead of a closed space, one thinks of an open space, then also the air may be referred to as being unsaturated or saturated with respect to water vapor, depending on the partial pressure of the water vapor in the air. However, this requires that there be no air *current* present.

8.22.8 Saturation pressure and superincumbent pressure

I have defined the boiling point under any given pressure as the temperature at which a liquid coexists with its vapor in equilibrium, the vapor being then a saturated one and the pressure being, by definition, the SVP at the boiling point. But when we boil water (or any other liquid) in an open vessel we do not observe any equilibrium between the water and the vapor above its surface since the air is usually *unsaturated* with respect to the water vapor. What, then, is it that makes the water boil?

Suppose that water is being boiled in an open vessel under an atmospheric pressure of $1.01 \times 10^5 \text{ Pa}$ (76 cm of Hg) and that the vapor pressure of water in the atmosphere is

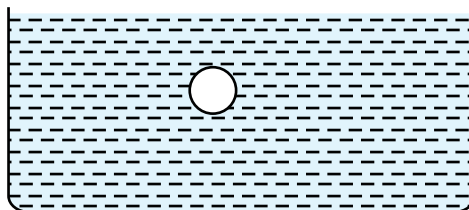


Figure 8-34: Formation of a bubble (magnified for the sake of illustration) in water in the process of boiling; the pressure inside the bubble being the same as the air pressure above the water surface, the bubble does not burst and rises to the top of the surface where it mixes into the water vapor in the air.

2.33kPa (1.75 cm of Hg). In this case, the water will boil when heated to a temperature of 373K, at which the SVP of water equals 101kPa, the atmospheric pressure over the water surface in the vessel, referred to as the *superincumbent pressure*. The air, however, *continues to remain unsaturated* with respect to the water vapor in it.

As the water is heated, bubbles made of saturated water vapor tend to form in it. At temperatures below 373K, the SVP of water is less than 101kPa, the superincumbent pressure, and so the bubbles collapse as soon as they are formed since the pressure inside the bubbles cannot compete with the pressure outside. At 373K, however, the two pressures become equal, and the bubbles containing saturated water vapor can grow in size, coexisting with the water (fig. 8-34). However, the bubbles cannot remain in mechanical equilibrium due to the buoyancy force on these, and they rise to the water surface where they get mixed with the water vapor in the atmosphere.

In other words, *water (or, for that matter, any other liquid) in an open vessel boils precisely at that temperature at which its SVP becomes equal to the superincumbent pressure (i.e., the pressure above the liquid surface)*. This is sometimes used as an alternative definition of the boiling point of a liquid.

The definition is, in principle, applicable to a liquid being boiled in a closed vessel as well, but in that case as soon as the liquid starts boiling, the addition of vapor above the surface leads to an increase in the superincumbent pressure which, in turn, requires a higher temperature for the liquid to boil. This, for instance, is the principle of cooking with a pressure-cooker where the water is made to boil at a temperature much higher

than its normal boiling point, resulting in a greater efficiency in cooking.

It may be mentioned here that, during the process of boiling, the bubbles are formed *throughout the volume* of the liquid being boiled, i.e., the change of state takes place everywhere throughout the region occupied by the liquid. This is a characteristic feature of all phase transitions where a homogeneous phase gets destabilized by the appearance of *nucleation centers* throughout its volume, these centers being made up of the new phase that is to appear in the transition.

The same feature characterizes the melting of a solid. During the melting, liquid drops are formed throughout the volume of the solid. When a piece of ice is heated, however, it is found to melt only at its surface, and no melting is observed in its interior. This is so because ice is a bad conductor of heat preventing heat being carried to its interior, and causing the melting to occur only at the surface where the heat is absorbed. In the melting of a metal, on the other hand, the melting process occurs in the entire volume of the metal.

8.22.9 Evaporation

In the above example of water in an open vessel, suppose that the water, instead of being heated to 373K, is kept at the same temperature as the atmosphere around it, say, 300K. The vapor pressure of water in the atmosphere being 2.33kPa, is less than the SVP of water at 300 K, which happens to be 4.45 kPa. This means that the air above the water surface is *unsaturated* with respect to water vapor and hence the water in the vessel cannot be in equilibrium with the water vapor in the atmosphere. What has to happen, then, is that water from the vessel will gradually get converted to vapor and will get mixed with the water vapor in the atmosphere. This process, referred to as *evaporation*, is distinct from boiling.

Under the circumstances referred to above, bubbles of saturated vapor cannot form and grow throughout the volume of the liquid. Instead, water molecules from the exposed surface of the water keep escaping into the atmosphere, mixing with the water vapor

in it. One similarity between the vapor molecules and the molecules in the liquid is that both are *mobile*, i.e., they can move around through rather long ranges and have their velocities distributed over a similarly wide range though, in a liquid, the motions of the various molecules are *correlated* to a greater degree compared to the correlations in a gas. This, in turn, is related to the fact that the average separation between the molecules is smaller in a liquid, with a correspondingly greater effect of the intermolecular interactions on the macroscopic properties of the material. Among the molecules of the exposed surface, the ones having a comparatively larger kinetic energy can overcome the pull of their neighboring molecules and escape into the atmosphere.

At the same time, the *reverse* process of vapor molecules from the atmosphere entering into the mass of the liquid also takes place (see fig. 8-35). However, when the vapor pressure in the atmosphere is less than the SVP of the liquid at the given temperature (as in the example we are now considering), this reverse process cannot quite match the forward rate at which the liquid molecules escape into the atmosphere, and the net result is that there is a comparatively slow depletion of liquid molecules from the exposed surface.

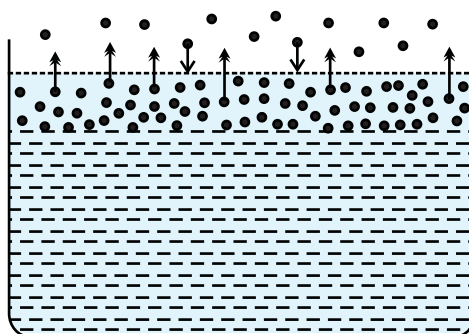


Figure 8-35: Two-way process (schematic) involving the escape of water molecules from the exposed surface to the atmosphere and the return of some molecules from the atmospheric vapor back into the mass of water, the latter being at a temperature lower than its boiling point under the given superincumbent pressure; the air above the water surface is assumed to be unsaturated, implying that the water cannot remain in equilibrium with the vapor; correspondingly, the rate of escape of water molecules into the vapor (double-headed arrows) is larger than the rate at which they return from the vapor to mix with the water molecules at the surface (single-headed arrows); water molecules in the liquid phase near the surface as also the ones in the vapor phase are represented with dots.

This process of depletion of liquid molecules from the exposed surface is precisely what has been referred to above as *evaporation*. If the vapor pressure above the liquid surface is close to the SVP at the given temperature, the rate of evaporation will be correspondingly slow since then the rate of return of molecules from the vapor into the liquid almost matches the rate of escape from the liquid into the vapor state. The evaporation stops completely as the vapor pressure above the liquid surface reaches the saturation value. This is observed when a sufficiently large mass of water is left to evaporate in a closed vessel and some water is found to remain while the rest evaporates. In the case of evaporation from an open vessel, on the other hand, this cannot happen since the molecules escaping into the vapor phase tend to get dispersed away. The dispersion of vapor molecules is aided if an air current blows over the surface, thereby accelerating the rate of evaporation.

Liquids with relatively high SVP's at ordinary temperatures evaporate quickly compared to ones with lower values of the SVP, and are termed *volatile*.

8.22.10 Relative humidity

The ratio of the vapor pressure of water in the atmosphere and the SVP of water at the atmospheric temperature is a quantity of considerable relevance, and is termed the *relative humidity* of the atmosphere.

$$(\text{relative humidity}) \ r = \frac{\text{partial pressure of water vapor}}{\text{SVP at given temperature}}. \quad (8-105)$$

It is commonly expressed in the form of a percentage figure.

Suppose that the partial pressure of water vapor in the air at a place is 25Torr (1Torr=1mm of Hg = 133Pa). Let the SVP of water at the atmospheric temperature be 33Torr. This then means that the relative humidity of air at that place is $(\frac{25}{33} \times 100\%)$, i.e., 76%. One can define the relative humidity in an alternative approach.

Suppose that a certain mass of air at a place is sealed off in a closed volume without altering its pressure, temperature, and composition. Let this mass of air contain ν mol

of water vapor. If one keeps on introducing more water vapor into the closed volume while keeping the temperature fixed, a stage will be reached when the air in the volume will become saturated with water vapor. Let, at saturation, the amount of water vapor in the said closed volume be ν_0 mol. Assuming that the temperature of the air is T and the volume sealed off is V , the partial pressure of water vapor in the air we started with is given by $\frac{\nu RT}{V}$, while the SVP is $\frac{\nu_0 RT}{V}$, where the air and water vapor have been assumed to follow the ideal gas equation of state. According to (8-105), the relative humidity then works out to $r = \frac{\nu}{\nu_0}$. Thus the alternative definition of relative humidity is

$$(\text{relative humidity}) \ r = \frac{\text{amount of water vapor in a closed space}}{\text{amount of vapor required for saturation at given temperature}}. \quad (8-106)$$

8.22.11 Dew Point

Suppose that the temperature of air at a certain place is T , and the partial pressure of water vapor in the air is p_W , and that the SVP of water at the temperature T is $p_S(T)$. Assuming that $p_W < p_S(T)$, the air will be unsaturated with respect to water vapor. Now suppose that the air is cooled gradually, keeping the pressure and composition unchanged. Since the proportional amount of water vapor and the pressure remains unchanged, the partial pressure of water vapor in the air remains unchanged at p_W , but the SVP keeps on decreasing due to the reduction in temperature. At some lowered value of the temperature, say T_0 , the SVP $p_S(T_0)$ will become equal to p_W , i.e., the air will become saturated at the temperature T_0 . This temperature T_0 is then referred to as the *dew point* of the air.

In other words, the dew point, with reference to given atmospheric conditions, is the temperature at which the SVP of water equals the partial pressure of water vapor in the air under the given conditions. If the temperature of air is made to decrease below the dew point, some amount of water vapor in the air will condense in the form of water drops.

It also follows from the definition of dew point that an alternative definition of relative

humidity is

$$(\text{relative humidity}) \ r = \frac{\text{SVP at dew point}}{\text{SVP under given atmospheric conditions}} \quad (8-107)$$

Problem 8-19

The SVP of water at 293K is $2.3 \times 10^3 \text{Pa}$. If the relative humidity at this temperature in a closed chamber of volume 5.0 m^3 be 60%, find the mass of water vapor present in the chamber, given that the molar mass of water is $18.0 \times 10^{-3} \text{ kg}\cdot\text{mol}^{-1}$.

Answer to Problem 8-19

SOLUTION: Making use of the definition of relative humidity, the partial pressure of water vapor in the chamber is $p = 2.3 \times 10^3 \times 0.6 \text{Pa}$. This must be equal to $\frac{\nu RT}{V}$, where ν stands for the amount in mole of water vapor in the chamber. Using the given values of T ($=300\text{K}$) and V ($=5.0\text{m}^3$), and substituting $R = 8.31 \text{ J}\cdot\text{mol}^{-1}\cdot\text{K}^{-1}$, one gets $\nu = 4.27$. Hence, the mass of water vapor present is $4.27 \times 18 \times 10^{-3} \text{kg}$, i.e., 0.077kg .

Problem 8-20

What is the change of entropy of one mole of a monatomic liquid, whose latent heat of evaporation is L , when it is converted into its saturated vapor at temperature T , and pressure p , the latter being the SVP of the liquid at temperature T . If p, p' be the SVP's at temperatures T, T' , estimate the difference in the molar entropies of the liquid when the pressure and temperature are changed from (p, T) to (p', T') .

Answer to Problem 8-20

HINT: Let the suffixes '1' and '2' be used to denote the liquid and gaseous phases. Considering the reversible process of conversion of one mole of the liquid to the gaseous phase at pressure and temperature (p, T) (such a process is possible since the liquid coexists with the gaseous phase in equilibrium, which need be disturbed only infinitesimally for the conversion to take place), the heat absorbed by the liquid is, by definition, the molar latent heat of evaporation, say, L . Since the temperature remains constant during the process, we have, $s_2 - s_1 = \frac{L}{T}$.

Considering the process of conversion at (p', T') , we similarly obtain $s'_2 - s'_1 = \frac{L}{T'}$ (strictly speaking, we should use L' instead of L here since the latent heat changes with transition temperature, but we take $L' \approx L$, which still gives us a good estimate). Thus, the change of entropy of the liquid for a change from (p, T) to (p', T') is $s'_1 - s_1 = s'_2 - s_2 - L(\frac{1}{T'} - \frac{1}{T})$. The change in entropy of the vapor for a change from (p, T) to (p', T') can be worked out by integration from the formula (8-47), by imagining a quasi-static process from the initial to the final state, and by making use of the ideal monatomic gas formulae $U = \frac{3}{2}RT$, $pV = RT$ (again, as an approximation), so as to obtain $s'_2 - s_2 = \frac{3}{2}R \ln \frac{T'}{T} + R \ln \frac{pT'}{p'T}$. This gives, finally, $s'_1 - s_1 = \frac{3}{2}R \ln \frac{T'}{T} + R \ln \frac{pT'}{p'T} - L(\frac{1}{T'} - \frac{1}{T})$.

8.23 Transmission of heat

The transmission of heat from one region to another in space can take place through one or more of three processes, namely, *conduction*, *convection*, and *radiation*. Conduction and convection require a medium through which the heat is transferred, while thermal radiation can take place in vacuum.

8.23.1 Conduction

When one end of a solid rod is heated, it is found that the other end also gets heated after an interval of time. The process by which heat is transmitted from one end of the rod to the other end in this instance is referred to as *conduction*. For some materials, conduction takes place at a rapid rate while, for some others, the process of conduction is found to be a relatively slower one. These are termed *good* and *bad* conductors of heat respectively. I will introduce below a physical characteristic of a material termed *thermal conductivity* (or, simply, *conductivity*). It is the numerical value of the conductivity of a material that determines whether it qualifies as a good or a bad conductor.

The mechanism underlying the process of conduction can be explained by referring to the *kinetic theory* of materials, according to which the atoms and molecules of a body are in perpetual random motion even when the body is in a state of equilibrium, and the mean energy of this motion increases with the temperature of the body. Such random motion of the microscopic constituents is sometimes referred to as *thermal motion*. If a certain part of a body (say, A) is at a relatively higher temperature then the molecules in this part have, on the average, a higher energy compared to those in other contiguous parts.

As a result of the interaction of the molecules in the region A with those in a contiguous region, say, B, energy is, on the average, transferred from the former to the latter. The part of the body occupying the region B is thereby heated up, and the process appears as a transfer of heat from the part A to the part B. The molecules in B, in turn, interact with those in another contiguous part, say C, and, on the average, transfer some of their energy to the latter. In this manner, heat is transferred from A to B, from B to C, and so

on to more distant parts of the body.

What is of special significance to note here is that, while transferring part of their energy to the molecules of a neighboring region, the molecules *do not get displaced from their own mean positions*, i.e., they are not themselves transferred from one part of the body to another. It is only a part of their energy that gets transferred by means of molecular interactions, commonly referred to as *collisions*, though the imagery of an *impact* does not necessarily apply to the interaction of the molecules. Interestingly, the interaction between the molecules of neighboring regions need not even be a direct one, and may be mediated by *free electrons* as in crystalline materials. Whatever be the mode of interaction, no part of the body under consideration moves from one region of space to another. Such movement, however, does occur in *convection*, as we will see later. Convection does not take place in solid bodies. Conduction, on the other hand, is relevant in the transfer of heat through a solid body, and it is the conductivity of the material making up the body that determines the rate with which the transfer takes place. Conduction can also take place in a fluid medium where a competing process of heat transfer is convection, the latter being, at times, the dominant mode of transfer.

In the case of thermal conduction in a liquid or a gas, it cannot be said that the mean positions of the molecules remain unchanged in conduction, since the molecules in a fluid are themselves in perpetual random motion. What happens here is that, energy is transferred by means of molecular collisions from any small region in the fluid to a contiguous region without the mean numbers of molecules in the two regions getting altered. What distinguishes the process of conduction in a liquid from that of a gas is the extent to which the mutual potential energy between the molecules of a small region gets transferred to that of a contiguous region, as compared to the kinetic energy.

Digression: Thermal and electrical conduction.

Materials that are good conductors of *electricity* are commonly found to be good conductors of *heat* as well. Metallic substances, in particular, are very good conductors of both electricity and heat. The process of conduction in these substances takes place

by means of *free electrons*. A notable consequence of the crystalline structure of these materials is that the outer electrons in the atoms in the crystal lattice get detached from the bondage of the respective nuclei and become *mobile* throughout the crystal lattice. These free electrons behave somewhat like the molecules of a gas, having random thermal motions of their own, the mean energy of the thermal motion increasing with the temperature of the body under consideration.

When a region in a conductor is heated, the mean energy of the atoms making up the crystal structure in that region increases, with an attendant increase of the mean energy of the free electrons in that region by virtue of interactions ('collisions') between the atoms and the electrons. Part of this increased energy of the free electrons then gets transferred to the electrons in a neighboring, colder region, eventually resulting in an increase of the mean energy of the atoms in that region, whereby the latter gets heated up. The process continues, with heat being conducted to more and more distant regions.

Thus, the 'gas' of free electrons, participating in the conduction process in a good thermal conductor, enhances the conductivity as compared to the conductivity of bad conductors, or *insulators* which lack in the pool of free electrons. It is the free electrons again that play a significant role in electrical conduction as well, as we will see in chapters 12 (see section 12.2) and 19.

Fig. 8-36 depicts schematically how heat is conducted from one end of a solid rod towards the other end through successive regions A, B, C, . . . of the rod without these parts of the rod getting displaced from their mean positions.

8.23.1.1 Thermal conductivity

Referring to the process of thermal conduction through a given material, imagine a thin layer of the material between a pair of parallel planes, say, A and B (fig. 8-37). Consider an area, say, δA on each plane, and suppose that the distance between the planes is δx . Assuming that the planes are at temperatures, say, T and $T + \delta T$, we ask the question: at what rate will heat flow from A to B by conduction? We assume here that the quantities

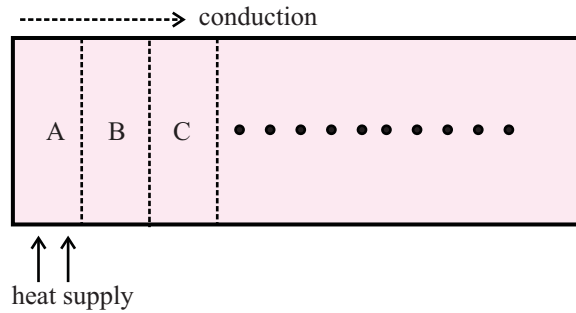


Figure 8-36: Conduction of heat through a rod (schematic); the mean energy of thermal motion of the atoms and molecules (as also of the *free electrons*) being relatively high at the heated end A, part of this energy is passed on to the neighbouring region B by means of collisions, thereby heating up this region; the process is repeated through other successive regions (B to C, and so on), constituting heat conduction; in the process the atoms and molecules in the respective regions of the rod do not, on the average, move away from their mean positions.

δA , δx , and δT to be of vanishingly small magnitudes. Experimental observations then tell us that the amount of heat conducted from A to B per unit time is proportional to the area δA and the temperature difference δT , and is *inversely* proportional to the thickness of the layer δx . This is again a quantity of vanishingly small magnitude which we denote by, say, δQ .

Thus, δQ can be expressed in the form

$$\delta Q = -K\delta A \frac{\delta T}{\delta x}, \quad (8-108)$$

where K is a constant for the material under consideration, and the negative sign is introduced in consideration of the fact that δQ can be positive only if δT is negative, i.e., heat will flow from A to B only if B is at a lower temperature than A, which means that the negative sign in the above formula actually makes the constant K a positive one.

In writing formula (8-108), we implicitly assume that the temperature varies along a direction perpendicular to the planes A and B, there being no variation in any direction parallel to these planes. In mathematical terms, this means that the *gradient* of temperature within the material is along the normal to the two planes (the temperature, as a function of the position vector within the material constitutes a *scalar field* and the flow of heat takes place parallel to

the *gradient* of this scalar field, in the opposite direction; refer to sec. 2.14.1).

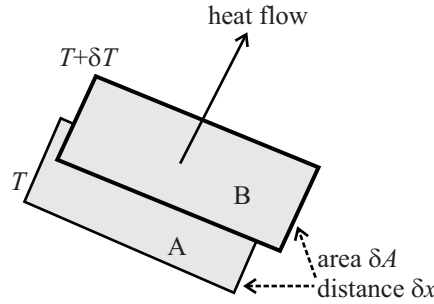


Figure 8-37: Illustrating the definition of thermal conductivity; heat flows through a given material from the section A to B in a direction perpendicular to both (direction of arrow); the conductivity determines the rate at which heat flows from A to B through the material; heat will actually flow from A to B only if δT is negative; in other words, the direction of heat flow is opposite to that of the temperature gradient.

This constant (K) is termed the *thermal conductivity* of the material under consideration. Its unit in the SI system is $\text{J}\cdot\text{m}^{-1}\cdot\text{s}^{-1}\cdot\text{K}^{-1}$.

While we have assumed δA , δx and δT to be small in the above definition, there may be certain set-ups where these need not be small. For instance, if the rate of change of temperature along the direction of heat flow is *uniform*, then the amount of heat conducted in unit time through a distance x between two parallel section A and B in the material, each of area, say, A , will be given by the expression

$$Q = -KA \frac{T_2 - T_1}{x}, \quad (8-109)$$

where T_1 and T_2 are the temperatures of the sections A and B respectively, and K is the thermal conductivity of the material under consideration.

1. Going over to the limit $\delta x \rightarrow 0$ in eq. (8-108), one arrives at the heat flow equation in one dimension: assuming the flow of heat to be along the x-axis of a co-ordinate system, the rate of flow of heat per unit area through a plane perpendicular to the

x-axis at any given point is given by

$$q = -K \frac{dT}{dx}, \quad (8-110)$$

where $\frac{dT}{dx}$ is referred to as the temperature gradient at the point under consideration, and q as the thermal current.

2. The question arises as to what determines the direction of flow of heat in the case of flow of heat in *three* dimensions? This depends on the temperature *distribution* in the material body under consideration, and has already been indicated above. The rate of heat flow is, in reality, a *vector* quantity. Given any point P with position vector, say, \mathbf{r} in the body, there corresponds a definite direction of flow of heat determined by the way the temperature varies around the point P. More precisely, one can talk of a vector, called the *temperature gradient* at P, whose Cartesian components are the three *partial derivatives* $\frac{\partial T}{\partial x}$, $\frac{\partial T}{\partial y}$, and $\frac{\partial T}{\partial z}$ of the temperature at the point P (see section 2.14.1 for a brief explanation of the concept of gradient of a scalar field). The flow of heat in a close neighborhood of P then occurs *in the direction opposite to that of the temperature gradient*. The vector representing the amount of heat flowing per unit area (imagined around P, with the normal to the area directed along the temperature gradient) per unit time is termed the *heat current* at P. The component of this vector along any given direction, say along the x-axis of a Cartesian system is then given by the expression $-K \frac{\partial T}{\partial x}$, which essentially expresses the same thing as eq. (8-110). All this is in elaboration of what has already been mentioned above.
3. While defining the thermal conductivity, I have assumed that the relation between the heat current and the temperature gradient is determined by a single constant K characterizing the material under consideration, or in other words, that the thermal conductivity is a *scalar* quantity. This is indeed so for conduction through liquids, gases, and *isotropic* solids, for which the heat current and the temperature gradient at any given point are parallel to each other (pointing in opposite directions). Certain crystalline substances, however, are *anisotropic* in the sense that the conduction along different directions in space are not all similar. For these materials, the relation between the heat current and the temperature gradient is determined not by a single scalar constant, but by a *set of* constants where it is seen that the heat current and the temperature gradient are

not, in general, parallel to each other.

8.23.1.2 Thermal diffusivity

Suppose that the thermal conductivity, density, and specific heat of a given material are respectively K , ρ , and s . Then the ratio $\frac{K}{\rho s}$ is termed the thermal *diffusivity* of the material. Denoting this by the symbol κ , one gets the mathematical expression

$$\kappa = \frac{K}{\rho s}. \quad (8-111)$$

According to this definition, the unit of diffusivity works out to $\text{m}^2 \cdot \text{s}^{-1}$. We will now look at the relevance of this new constant in the context of thermal conduction in a material.

8.23.1.3 Stationary and non-stationary heat flow

Suppose that a rod, initially at a uniform temperature (which we assume to be equal to the temperature of the surrounding air), is being heated at a constant rate at one end (A in fig. 8-38). Thermometers inserted at various points along the length of the rod (points B, C, D, etc., in the figure) can be used to measure the temperatures at these points at various different time instants. How will the readings of these thermometers be found to change with time?

In this case, heat will be conducted away from the end A and get transferred toward the other end of the rod. The reading of the thermometer at B will first be found to register an increase in temperature, subsequent to which, the readings of C, D, etc., will also rise in succession. This stage during which the temperatures at all the points on the rod keep on changing with time, is termed a *non-stationary* state in thermal conduction.

As the readings of the thermometers are checked at regular intervals, it will ultimately be seen that each thermometer attains a certain reading of its own which then remains constant in time. This constant reading will have the maximum value for the thermometer at B, and will have decreasing values for the successive thermometers inserted at more distant points (C, D, etc.,) away from A. This stage of thermal conduction where each point attains a steady temperature depending on its location on the rod is termed

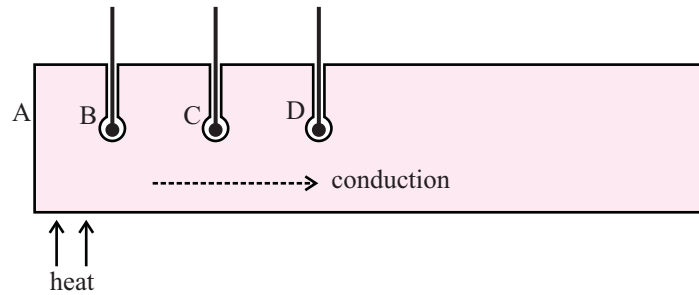


Figure 8-38: Set-up to illustrate stationary and non-stationary states in thermal conduction; heat is supplied to a rod, initially at a uniform temperature, at one end and flows along the length of the rod towards the other end; thermometers are inserted at various points like B, C, D, along the length of the rod; readings of these thermometers taken at regular intervals of time show, at first, a non-stationary condition in which each thermometer registers an increase in its readings; eventually a stationary condition sets in when each of the thermometers shows a constant reading of its own.

a *stationary* or *steady* state.

A mathematical analysis shows that the rate of rise of the temperature in the non-stationary state at any given point of the rod is determined precisely by the diffusivity κ defined in eq. (8-111). The temperature distribution along the length of the rod in the steady state, on the other hand, does not depend on the diffusivity and is determined by the conductivity K , in addition to a number of other relevant parameters.

What happens here can be explained as follows.

Consider a thin layer of the material between sections, say, S_1 and S_2 of the rod (not shown in fig. 8-38), where heat flow by conduction takes place from the former to the latter, coming into the layer from the region of the rod lying to the left of S_1 and moving out to the region lying to the right of S_2 . The *difference* between the amount of heat entering into the layer under consideration through S_1 and that leaving through S_2 in a unit time interval gives the *net* rate of supply of heat into the layer. Calling it q , the energy accounting for the layer can be expressed by an equation of the form

$$q = q_1 + q_2, \quad (8-112)$$

where q_1 is the rate at which heat flows out from the layer into surrounding regions, principally by convection and radiation, and q_2 is the rate at which heat accumu-

lates in the layer, causing its temperature to change. Since the rod is initially at a uniform temperature, equal to the temperature of the surrounding air, the rate of heat lost to the surroundings by convection and radiation remains small (see sections 8.23.2, 8.23.3) in the initial stages of conduction, and one then has $q \approx q_2$. Denoting the rise in temperature of the layer under consideration per unit time by ΔT , one has, $q_2 = \rho V s \Delta T$, where V denotes the volume of the layer. Recalling that rate of heat flow by conduction involves the thermal conductivity K , one concludes that the rate of increase of temperature of a thin layer in the initial, non-stationary state is governed by the constant $\frac{K}{\rho s}$, i.e., by the thermal diffusivity of the material.

Late in the process, however, when the temperatures at various points of the rod become different from one another and from the surrounding air, q_1 increases and eventually becomes close to q , when q_2 diminishes to a small value, $q_2 \approx 0$, when the rates of change of the temperatures tend to zero. This is the steady state when the entire heat supplied at the end A in any given time gets transferred to the surroundings by convection and radiation. In this stage, the temperature at a point close to A (say, at B in fig. 8-38) reaches a steady value higher than that at a more distant point (like, say, C). The net amount of heat entering into any part of the rod at this stage by conduction gets lost from that part to the surroundings by convection and radiation.

Problem 8-21

A composite rod is made up of two rods of different materials but of identical cross-sections, each of area A . The rods are of lengths l_1 and l_2 , and are made of materials of thermal conductivities K_1 and K_2 . One end of the composite rod is maintained at a temperature T_1 , while the other end is kept at a lower temperature T_2 . What will be the temperature at the junction of the two rods in the steady state?

Answer to Problem 8-21

HINT: In the steady state the rates of heat flowing through the two rods must be the same. Hence, if the temperature at the junction is T , then we must have

$$K_1 A \frac{T_1 - T}{l_1} = K_2 A \frac{T - T_2}{l_2},$$

giving

$$T = \frac{\frac{K_1}{l_1}T_1 + \frac{K_2}{l_2}T_2}{\frac{K_1}{l_1} + \frac{K_2}{l_2}}$$

8.23.2 Convection

The principal means of heat transfer in liquids and gases under commonly observed situations is *convection*. A major feature of convection distinguishing it from conduction is that, in convection, there occurs a continual *displacement* of constituent parts of the liquid or gas from one region to another, where the interchange of positions of these parts accounts for the transfer of heat.

Suppose that heat is being supplied from below to a liquid contained in a vessel (see fig. 8-39). As a consequence of the heat supply, the liquid near the bottom of the vessel gets heated first and its density decreases (an exception to this rule is provided by water in the temperature range from 0° C to 4° C). This lighter liquid rises up due the force of buoyancy, displacing the comparatively denser and cooler liquid from the upper layers, whereas the latter gets displaced towards the bottom of the vessel. A convection *current* is thus generated in the vessel due to the interchange of positions of the different parts of the liquid.

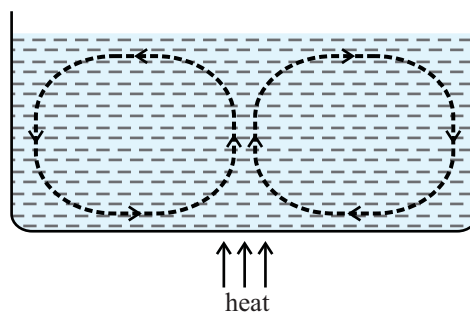


Figure 8-39: Convection in a liquid heated from below; the convection current has been indicated with arrows.

The liquid descending from the upper layers in turn gets heated due to the heat supplied from below the bottom of the vessel, and eventually rises up again as its density

decreases below the liquid now occupying the upper layers. The process continues, causing heat to be carried from bottom upwards, with an increase in the average temperature of the liquid in the vessel.

Along with convection, some amount of heat transfer takes place in a liquid or a gas by conduction as well. The molecules in one region in the fluid transfer part of their thermal energy to those in a neighboring region through molecular collisions without themselves getting displaced in the process from one region to another. In some situations, there occurs a competition between the two processes of conduction and convection and, at times, even a *turn-around* can take place with a change in the relative importance of the two processes in heat transfer.

For instance, in the above example of a liquid being heated from below, the dominant mode of heat transfer from bottom upward in the early stages of heating may be that of conduction. This is because of the fact that the temperature difference between the bottom and top layers being small in this early stage, the buoyancy force on the bottom layer may not be sufficient to overcome the *viscous* resistance to motion, and the convection current may fail to develop. However, later in the heating process, the temperature difference between the lower and upper layers increases and the density difference between these layers becomes sufficient to set up a convection current, overcoming the viscous resistance.

Convection currents have significant roles to play in various situations of practical interest and in numerous natural phenomena. In order to keep a room warm in a cold climate, heat is generated in a fireplace or by means of an electric heater near a bottom corner of the room. The air in the immediate neighborhood of the fireplace or the heater gets heated and becomes lighter, initiating a convection current and eventually causing an increase in the average temperature in the room.

The air currents set up in the atmosphere around us are, in reality, often caused by convection. Well known instances of these are the land-breeze and the sea-breeze blowing near large water masses such as lakes or seas. Such convective air currents are

important in the maintenance of climatic temperateness as also in initiating weather and climatic changes. Numerous convection currents are also found at various levels of sea water. Convection currents in a blast furnace are controlled by varying the amount of fuel in the furnace and by varying the intensity of blast (which causes a change in the rate of burning of fuel as well), as a result of which the temperatures at various heights in the furnace are maintained at desired levels.

8.23.2.1 Natural and forced convection

As we have seen above, a convection current is generated in a liquid or a gas in the process of heat transfer by convection, whereby a *flow* is set up involving the displacement of various parts of the fluid. A flow can be set up in a fluid by other means as well. For instance, if a pressure difference is maintained between two regions in a fluid, then a flow takes place from the region of higher to one of lower pressure.

If such flows, driven by causes other than a temperature difference, are made to occur in the fluid, then these may affect the rate of thermal convection in it. The term *forced convection* is used to refer to a process of convective heat transfer modified by such flows generated by other means. If, on the other hand, no such effects caused by pressure differences or other means are brought in to modify the convective current, then the process of heat transfer is referred to as *natural* convection.

8.23.3 Thermal radiation

When we warm our hands in a cold day by stretching these above an oven or a hearth, the hands feel warm mainly due to the process of convection in the air above the oven. However, a second process, that of *radiative heat transfer* also comes into play in this case, which becomes more apparent when, instead of the hands being placed above the oven, these are held at some distance near the bottom of it. Once again the hands feel warm, but now the role of convection in the heat transfer is considerably less. What happens now is that some heat gets spread out in space on all sides around the oven by the process of *radiation*, which occurs independently of conduction and convection.

The origin of thermal radiation lies in processes of a microscopic nature occurring within radiating bodies. All matter is composed of atoms and molecules, in which *electrons* move about in a number of stationary *orbitals* (see chapter 18 for an introduction to basic ideas in atomic physics), occasionally making *transitions* from one orbital to another. Usually an electron tends to reside in the orbital of lowest energy, referred to as its *ground state* while it may occasionally receive a supply of energy from other sources so as to move up to an orbital at a higher energy level, termed an *excited state* of the electron. The electron, however, does not stay in the excited state for ever, and quickly gets back to a state of lower energy, preferably the ground state. In this process of *de-excitation* the electron releases some amount of energy into the surrounding space, causing variations in the *electric and magnetic* field intensities at various points therein. These fluctuations in the electric and magnetic field intensities are generated because of the fact that the electrons are *charged* particles that can act as sources of electric and magnetic fields (see chapters 14 and 16 for the basic ideas involved here).

Along with the radiation from such *bound* electrons, radiation of energy can also occur from electrons not bound to specific atoms or molecules. These electrons, in course of their motion, get *scattered* from the atoms and molecules of the material in which they move. Such scattering events correspond to acceleration and deceleration of charged particles, and once again give rise to fluctuations in the electric and magnetic field intensities in regions close to the source.

The fluctuations in the electric and magnetic field intensities initiated in regions close to the radiating source, are carried to more distant regions in the form of *electromagnetic waves*. In the process, an amount of *energy* is transmitted from regions close to the source to the more distant regions. Where does this energy come from in the first place? It is precisely the energy that the electrons inside the radiating body release in the process of de-excitation from higher energy orbitals to lower energy ones or in the process of getting scattered.

So, here is a picture of what happens in thermal radiation. When a body is heated up so as to serve as a source of thermal radiation, the energy of random motion of its

atoms and molecules increases and part of this energy gets transferred to the electrons inside the atoms, or to the free (i.e., mobile) electrons, by means of collisions. On being endowed with this increased energy, the electrons make transitions to excited states, being again de-excited to states of lower energy. In the process, they release a part of their energy into the surrounding space which appears as the energy of electromagnetic fluctuations in the vicinity of the source. Added to this is the radiation generated in the scattering events of electrons not bound to specific atoms or molecules, where these events increase in frequency of occurrence as the temperature of the source is made to increase.

This radiated energy, in turn, spreads out to more distant parts of space in the form of electromagnetic waves. When an electromagnetic wave is incident on a second body, part of the energy carried by the wave gets absorbed by this second body, which thereby gets heated up. This, precisely, is the process of radiative transfer of heat from the first body, the source, to the second body.

The picture remains somewhat incomplete unless one explains how and why the energy gets transferred from the electromagnetic wave to the recipient body. The explanation lies in the fact that just as a charged particle in motion is capable of generating an electric and a magnetic field, causing an electromagnetic wave to be set up, conversely, an electromagnetic field can alter the state of motion of a charged particle by way of exerting a force and imparting energy to it. It is this action of the electromagnetic wave on the electrons in the recipient body that raises these electrons to states of higher energy of their own, and this added energy finally appears as the energy of random thermal motion of the atoms and molecules through the mechanism of collisions.

One major characteristic of electromagnetic waves is that these can travel through vacuum. Indeed, charged particles in motion can set up electric and magnetic fields even when not located in a material medium. The speed with which an electromagnetic wave propagates through vacuum is $3 \times 10^8 \text{ m}\cdot\text{s}^{-1}$, commonly referred to as the *velocity of light*.

While the speed of an electromagnetic wave in vacuum is a universal characteristic of electromagnetic waves in general, the *frequency* of the wave is another matter altogether. Indeed, the electromagnetic radiation from a heated body is made up of a large number of *monochromatic* waves, each monochromatic wave having a specific frequency of its own. The frequencies making up the radiation from a heated body generally range from very low to very high values, making up what is known as the electromagnetic spectrum (see section 14.4.5). The total energy carried by the electromagnetic radiation is distributed among the various frequencies in a manner depending on the temperature of the radiating body. At low temperatures, relatively low frequencies predominate in the emitted radiation while, at higher temperatures, higher frequencies carry a relatively larger proportion of energy. At temperatures of around 500 K and above, a considerable proportion of energy is carried by frequencies belonging to a certain range referred to as the infra-red and microwave parts of the electromagnetic spectrum. Waves in this frequency range have the special ability to transfer their energy efficiently to other bodies on which they are made to be incident, *heating up* these bodies. The transfer of heat by radiation from one body to another takes place mostly through these infra-red waves and microwaves, sometimes jointly referred to as *thermal waves*.

When a radiating body is raised to a very high temperature, say, 3000 K or above, the radiation emitted by it contains quite a bit of relatively high frequency components, including those we perceive as *visible light*. The latter is made up of components causing the sensation of colors ranging from the red to the violet. At sufficiently high temperatures of the radiating body, all these components are present in the radiation, apart from the ones we have referred to as the thermal waves, and the body appears *white hot*.

However, among all the frequency components radiated from a white hot body, only those corresponding to infra-red waves and microwaves are capable of efficiently heating up a recipient body. In the process of emission and absorption of these components (corresponding what is commonly referred to as thermal radiation), the processes of scattering of mobile electrons commonly dominates over those involving transitions of bound electrons. Thermal radiation also involves changes in rotational and vibrational

states of molecules.

8.23.3.1 Stefan's law of radiation

The radiation coming out from a heated body is made up of *photons*, the latter being energy quanta of the electromagnetic field associated with the radiation (see section 16.9 for a brief introduction to the idea of photons). Corresponding to the energy of the radiation being distributed between monochromatic waves of various frequencies, the photons are also characterized by frequencies distributed throughout the electromagnetic spectrum. If the photons, which behave somewhat like gas molecules, form a system in thermodynamic equilibrium at any given temperature T , then the resulting electromagnetic field is referred to as *black body radiation* at temperature T . The most distinguishing feature of black-body radiation is the way the energy of the electromagnetic field is distributed among the various frequency components. This is referred to as the *black body distribution* or *Planck distribution* (see, once again, section 16.9).

The radiative transfer of heat from one body to another can be interpreted as a stream of photons emitted from the source hitting the recipient body, which is simply another way of saying that an electromagnetic disturbance reaches out from the former to the latter. These photons are, in a sense, like photons leaking out from a chamber containing black body radiation at the temperature of the source, and are thus characterized by the Planck distribution formula (you will find this formula written out in section 16.9).

The Planck formula, in turn, leads to a neat little expression for the *total* energy radiated from the source per unit time per unit surface area of the source. This goes by the name of *Stefan's law of radiation* and reads

$$H = \sigma T^4. \quad (8-113a)$$

In this equation, σ is a constant independent of the material the source is made of, and is known as the *Stefan constant*, its value being $5.67 \times 10^{-8} \text{ W}\cdot\text{m}^{-2}\cdot\text{K}^{-4}$, and T stands for the temperature of the source. Strictly speaking, this formula for the total energy radiated per unit time per unit emitting area (H) at temperature T is applicable to

what is known as a *black body*, which is defined to be a body with idealized properties, capable of absorbing completely whatever radiant energy is incident on it regardless of the frequency of the wave the energy is carried by. In reality, the surface properties of the emitting body depart somewhat from those of an ideal absorbing body, and the rate of total emission at any given temperature gets modified accordingly (indeed, the surface characteristics relating to the degrees of emission and absorption are related to each other, as we will see below in the context of *Kirchoff's principle*).

For an emitting body, then, which is not an ideally 'black' one, formula (8-113a) has to be modified to

$$H = \epsilon \sigma T^4. \quad (8-113b)$$

where ϵ is a constant characterizing the emissive property of the emitting surface, lying in the range $0 < \epsilon < 1$. It signifies the extent to which the total emission from the surface under consideration falls below that of an ideal black surface at the given temperature. Recall that the photons involved in the radiation from a heated body can be looked upon as those leaking out from a chamber filled up with black body radiation. According to this interpretation, the constant ϵ , referred to as the *emissivity* of the surface under consideration, can be looked upon as some kind of a transparency factor characterizing the surface, and effectively corresponds to the fraction of photons, incident on the surface either from the interior or from the exterior, that can pass through on to the other side. In general, the emissivity of a dark and rough surfaces is greater than that of a light colored and smooth one. Since a *black body* is one that completely absorbs radiation of all wavelengths incident on it, its surface has to be perfectly transparent to radiation of all wavelengths, and hence the effective transparency, the *emissivity*, is unity.

What is meant here by the term transparency is the efficiency with which a photon incident on the surface, say, from the exterior region, is passed on to the interior of the body. The same efficiency also characterizes the passage of a photon, incident from the interior, to the exterior region. A value less than unity for the transparency implies that there is some probability that a photon, incident from either side, is sent

back to where it came from.

8.23.3.2 Energy exchange in radiative transfer

If a body at any given temperature T emits energy by radiation in accordance with formula (8-113a) or (8-113b), then all bodies should give out heat and, in the process, get cooled. However, as well as giving out heat, a body also *receives* radiant energy incident on it, originating from other bodies, a part of which is absorbed by it. This corresponds to a heating up of the body, which competes with the process of cooling down, and the net result then depends on which of the two processes dominates over the other.

Fig. 8-40 depicts a body at a temperature T kept within a closed chamber, the latter being maintained at a temperature T' . In this case, the total energy radiated per unit time by the body into its surrounding region will be $\epsilon\sigma AT^4$, where A stands for the surface area of the body. Some amount of energy will also be incident on the body on being radiated by the enclosure, wherein a part will be absorbed by it. The energy so absorbed by the body is found to work out to $\epsilon\sigma AT'^4$.

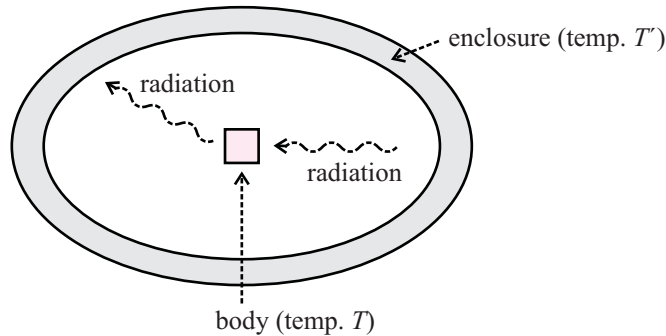


Figure 8-40: Radiative energy exchange between a body and the enclosure it is kept in.

This means that if the temperature of the body is greater than that of the enclosure ($T > T'$), then the former will, on the whole, lose energy by radiation and will thereby get cooled by radiative heat transfer to the enclosure. If, on the other hand, the enclosure is warmer ($T' > T$), radiative heat transfer will occur in the opposite direction, from the

enclosure to the body. The net rate of radiation from the body is then

$$H = \epsilon\sigma A(T^4 - T'^4). \quad (8-114)$$

One can generalize these considerations to a situation where there are a number of radiating bodies, each at a given temperature of its own, the net rate of heating or cooling of any one of these bodies being determined by mutual radiative transfers between this body and all the other bodies in the system.

As a rule, the process of energy exchange by radiative heat transfer between two bodies results in the hotter body getting cooled and the colder getting heated up. The medium through which the electromagnetic waves propagate, carrying energy from one body to the other (in some situations this energy can be carried even through vacuum) can be looked upon as a means of establishing thermal contact between the two bodies.

Electromagnetic waves carrying radiant energy can propagate through solid, liquid, and gaseous media, where the speed of propagation depends on the material of the medium. In most solids and liquids, however, the electromagnetic waves die down quickly, being absorbed by the material itself. In other words, radiative heat transfer does not occur efficiently through a dense medium. In a gas at a low pressure, on the other hand, electromagnetic waves carrying radiant energy can propagate through long distances.

Problem 8-22

A block of metal in the form of a cube of edge length 0.15 m is heated to a temperature of 700 K. If the emissivity is 0.50, find the rate of radiation of heat from the cube. If the temperature of the surroundings be 300 K, what is the net rate of loss of energy from the body?

Answer to Problem 8-22

HINT: The surface area of the cube is $A = 6 \times (0.15)^2 \text{ m}^2$. The rate of radiation is then $0.5 \times \sigma \times A \times (700)^4 \text{ J}\cdot\text{s}^{-1}$, ($\sigma = 5.67 \times 10^{-8} \text{ J}\cdot\text{m}^{-2}\text{s}^{-1}\text{K}^{-4}$), which works out to $9.2 \times 10^2 \text{ J}\cdot\text{s}^{-1}$ (approx). The net rate of energy loss from the body is $0.5 \times 5.67 \times 10^{-8} \times 6 \times (0.15)^2 \times (700^4 - 300^4) \text{ J}\cdot\text{s}^{-1}$, i.e., $8.9 \times 10^2 \text{ J}\cdot\text{s}^{-1}$ (approx).

Problem 8-23

One end of a rod of length $l = 0.3$ m is maintained at a temperature of $T_0 = 50$ K while the other end is blackened and exposed to surroundings at a temperature of $T = 500$ K. If the rod be made of a good thermal conductor of conductivity $K = 1500 \text{ J}\cdot\text{s}^{-1}\cdot\text{m}^{-1}\cdot\text{K}^{-1}$, and if there is no loss of heat except from the end faces of the rod, calculate the temperature of the exposed end.

Answer to Problem 8-23

HINT: The rate of absorption of radiant heat at the exposed end (say, A) is to be the same as the rate of conduction toward the other end (say, B). Since the rod is made of a good conductor, one can assume that the required temperature T' at the end A is only slightly higher than T_0 (for the sake of consistency, this is to be checked against the result obtained), i.e., $T' = T_0 + x$, where $x \ll T_0$. If A be the area of cross-section of the rod then one has $\sigma A(T^4 - T_0^4) \approx K \frac{A}{l} x$ (σ = Stefan's constant), where, in the left hand side we have replaced $(T_0 + x)^4$ with T_0^4 . One can go even further and neglect T_0^4 compared to T^4 (approximation is the name of the game!, though the game is always to be played with cool reasoning). Equating this rate of absorption at the blackened end with the rate of conduction, i.e., to $\frac{KAx}{l}$ and using given values, one finds $x = 0.71$ K.

One can have a better approximation for x , if necessary, by using $\sigma A(T^4 - (T_0 + 0.71)^4)$ on the left hand side of the above equation.

8.23.3.3 Kirchhoff's principle

When radiant energy carried by electromagnetic waves of any given wavelength is incident on the boundary surface of a body, a part of it is sent back to the medium where it came from, while the remaining is transmitted to the interior of the body. A fraction of the radiant energy entering into the body, depending on the composition and dimensions of the latter, penetrates through it and is eventually released to its surroundings while the remaining portion is *absorbed* into the body. For our present purpose, however, this distinction between the energy penetrating through the body and that retained in it is not essential, and whatever part of the energy incident on the surface enters into its interior will be said to be absorbed into it.

Referring to the radiation incident on the boundary surface of the body, the relative amounts of energy entering into it (i.e., ‘absorbed’ by it according to our present terminology) and that sent back into the surrounding medium depend on the wavelength of the radiation and the nature of the surface. A quantitative expression of the degree of absorption at any given wavelength can be given in terms of a coefficient termed the *absorption coefficient* of the surface for the wavelength under consideration.

Along with the absorption of radiant energy incident on a body, there occurs *emission* from it as well, where the degree of emission can be expressed quantitatively in terms of an *emission coefficient* depending on the wavelength and the nature of the emitting surface. Kirchhoff’s law of emission and absorption then states that, for a given wavelength, the absorption and emission coefficients of various surfaces follow a relation of *proportionality* to each other: *good absorbers are also good emitters*.

The absorption coefficient for a perfectly black body is unity for all wavelengths while its emission coefficient can be mathematically worked out from Planck’s formula. Considering any other body, its absorption coefficient at any given wavelength will be less than unity by a certain factor that can be identified as the ‘transparency factor’ mentioned above, now pertaining to the wavelength under consideration (in contrast, the transparency factor occurring in formula (8-114) is in the nature of an average over all wavelengths). According to Kirchhoff’s law, the emission coefficient of the body at that wavelength will be reduced by the *same* factor relative to the emission coefficient of a black body. Evidently, the emission coefficient of a black body for any given wavelength is larger than that for any other body.

While the radiation emitted by a heated body is made up of *all* frequencies ranging *continuously* from zero to infinitely large values, one also observes radiation at certain *special* frequencies depending on the composition of the body and the physical condition it is in. For instance, when hydrogen gas is heated to high temperatures, it emits radiation at a set of frequencies characteristic of hydrogen, differing from the frequencies emitted by other elements. On the other hand, when radiation made up of a continuous range of frequencies is passed through cold hydrogen gas, certain components

of the radiation are *absorbed*. Kirchhoff's principle, stated above, holds in respect of the emission and absorption of these special or characteristic frequencies as it does for other frequencies as well: a substance absorbs those characteristic frequencies which it emits. In other words, emission and absorption go hand in hand. However, not all of the characteristic emission frequencies are absorbed with equal efficiency by a material.

8.23.3.4 Newton's law of cooling

Think of a heated body kept in the air. Heat is lost by the body to the surrounding air and eventually to other systems as well, by means of convection and radiation. This results in a decrease of the internal energy of the body and thereby a decrease in temperature (we assume that no internal energy is being supplied to the body in the form of work). In this instance, the cooling of the body occurs principally through convection and radiation. While conduction may also be responsible for the cooling in some situations, we will leave it out of our considerations now.

A useful formula for the rate of cooling of a body due to convection is referred to as *Newton's law of cooling*. Suppose that the temperature of a body is T , and that of the air surrounding it is T_0 , i.e., the temperature difference between the body and its surroundings is $T - T_0$. According to Newton's law of cooling, the rate of decrease of temperature of the body due to convection is *proportional to this temperature difference*, i.e., to $T - T_0$. In other words, one has

$$\frac{dT}{dt} = -A(T - T_0), \quad (8-115)$$

where A is a constant independent of the temperature of the body and of the surroundings, but depending on factors like the area of the exposed surface of the body, the viscosity of the surrounding medium (we have assumed the medium to be air, but Newton's law of cooling remains valid for other fluid media as well), and the rate of flow, if any, of the medium past the body. The negative sign in the above formula indicates that there occurs a cooling only if the body is at a temperature higher than that of its surrounding medium. It is assumed here that the body under consideration does not gain or lose heat by exchange with other bodies.

It has been found experimentally that formula (8-115), expressing Newton's law of cooling, is a useful one explaining the cooling of bodies if the temperature difference $T - T_0$ is not too large. It is, however, not an exact formula that can be derived from fundamental principles. In some circumstances, a more general formula of the following form gives a better description of the cooling process

$$\frac{dT}{dt} = -A(T - T_0)^\gamma. \quad (8-116)$$

Here γ is another empirical constant of the order of unity, the value $\gamma = 1$ corresponding to the commonly stated form of Newton's law, i.e., eq. (8-115). While eq. (8-115) is applicable, in an approximate sense, to a range of situations involving forced convection, a formula of the form (8-116), with γ slightly larger than unity, works better for natural convection.

A formula similar to eq. (8-115) also applies to heat loss by *radiation* from a heated body to its surroundings, provided the temperature difference $T - T_0$ is small. However, this is again only an approximate rule, while a more exact formula can be derived from Stefan's law of radiation.

At times, a heated body is provided with an insulating coating to prevent heat loss to the surroundings from its exposed surface by convection. Heat is lost then from the outer surface of the coating by convection, while the direct process of heat loss from the body is one of conduction through the layer of insulation. In the steady state, the rate of heat loss by conduction, which is given by an expression of the form (8-115) (with A depending on the conductivity of the insulating layer and T_0 denoting the temperature of the outer surface of the layer) has to equal the rate of heat lost from the outer surface of the layer by convection and radiation, where the latter is again of the form (8-115) (with a different constant A , and with $T - T_0$ replaced with $T_0 - T'$, where T' denotes the temperature of the surrounding air).

Problem 8-24

Heat is supplied to a boiler at the constant rate $q = 1000\text{J}\cdot\text{s}^{-1}$. The thermal capacity of the boiler along with its contents is $s = 800\text{J}\cdot\text{K}^{-1}$, and its temperature T rises by $1.0\text{K}\cdot\text{s}^{-1}$ at an instant when $T = 330\text{ K}$. Assuming that the loss of heat from the boiler occurs in accordance with Newton's law of cooling and that the surroundings are at a temperature $T_0 = 300\text{ K}$, find the rate of rise of temperature of the boiler and its contents when $T = 360\text{ K}$.

Answer to Problem 8-24

HINT: Since the rate of supply of heat to the boiler has to equal the rate at which heat is absorbed by the boiler and its contents (i.e., $s\frac{dT}{dt}$) plus the rate of loss of heat, i.e., $A(T - T_0)$, where A is an appropriate constant (Newton's law of cooling), one has $q = s\frac{dT}{dt} + A(T - T_0)$. In the first instance, then, $1000 = 800 \times 1.0 + A \times 30$, i.e., $A = \frac{20}{3}\text{ J}\cdot\text{s}^{-1}\cdot\text{K}^{-1}$. With this value of A one obtains, for $T = 360\text{ K}$, $\frac{dT}{dt} = \frac{1000 - \frac{20}{3} \times 60}{800}\text{K}\cdot\text{s}^{-1}$, i.e., $0.75\text{K}\cdot\text{s}^{-1}$.

8.23.3.5 The greenhouse effect

Think of a body kept inside a closed chamber with walls made of glass. Suppose that thermal radiation emitted from a heated body at a high temperature kept outside the chamber penetrates the glass walls and is incident on the body within the chamber. One part of this incident radiant energy gets absorbed by the body, thereby heating it up. As we know, the body itself acts a source of radiation, releasing radiant energy in the form of electromagnetic waves.

The source of radiation outside the chamber being at a high temperature, this radiation is relatively rich in high frequency components (a consequence of Planck's distribution formula) and has the ability to penetrate thick layers of glass and numerous other materials without being absorbed to any appreciable extent. The temperature of the body within the chamber, on the other hand, is comparatively low, and the radiation from *this* body is richer in relatively *low frequency* components. These components of lower frequencies are less capable of penetrating through solid materials like glass. Consequently, a considerable part of the radiation from the body kept within the enclosure fails to come out of it, remaining trapped in the form of *stationary waves*, much like the waves in a microwave oven used for the purpose of cooking.

This one-way traffic of radiant energy from the source outside the chamber to its interior continues, gradually heating up the interior. One refers to this process as the *greenhouse effect* (fig. 8-41).

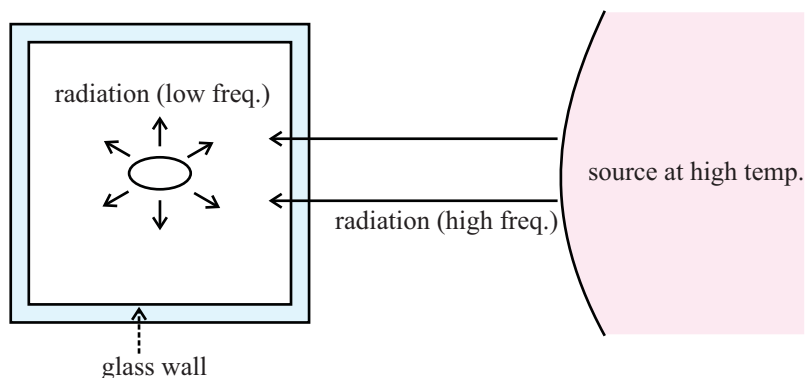


Figure 8-41: Illustrating the principle underlying the greenhouse effect.

This effect is made use of in nurturing green plants in a glass chamber during the winter, where sun-rays entering into the chamber keep the plants warm, since the radiation emitted by the plants inside the greenhouse cannot escape out of it.

In recent times, the average temperature of the earth and the atmospheric layer surrounding it has been found to increase at a rather steep rate. This phenomenon can also be traced to the greenhouse effect. Uncontrolled industrialization and urbanization has led to a notable rise in the proportional amount of compounds like carbon dioxide and chlorofluorocarbon (CFC) in the atmosphere. The role of the enclosure walls is played mainly by these two compounds in the atmosphere, along with water vapor. Thus, these gases allow radiant energy from the sun, rich in high frequency components, to pass through and heat up the earth and the surrounding layers of the atmosphere, but prevent the relatively lower frequency radiation emitted from these from being sent back to the outer space. The resulting trapped radiation, rich in infra-red waves, heats up the earth and the adjacent atmospheric layer.

Indeed, global warming is considered to be a major problem threatening human existence to-day. Moreover, this global warming is quite likely to be self-reinforcing since it

will result in the release of a large quantity of methane gas, trapped in arctic ice, into the atmosphere as the ice keeps on melting. Methane in the atmosphere is a much more potent greenhouse agent as compared to carbon dioxide and CFC.

It may, however, be mentioned that, according to an alternative point of view, the question of global warming needs reconsideration. For instance, it appears likely that global warming is correlated with *solar activity*, i.e., alternating cycles of increased radiation from the sun, and that the rate of global warming itself may undergo cyclical changes over long periods of time. At any rate, global climatic change is a complex process, yet to be adequately understood in its long term trends.

8.24 Supplement: Random variables and probability distributions

In mathematics and the physical sciences we come across numerous *constants* and, additionally, various different *variables*. For instance, the number $\sqrt{3}$, or the velocity of light in vacuum ($c = 3 \times 10^8 \text{ m}\cdot\text{s}^{-1}$) are constants. In contrast, a variable can undergo a change from one value to another, depending on the context. For instance, the temperature (T) of water kept in a vessel can change from time to time as a result of exchange of heat with other systems. Or, the length (l) of a simple pendulum may be made to change from time to time.

Variables may be classified into two broad groups, namely, *deterministic* variables, and *random* ones. For instance, while the variable l in a pendulum experiment can have any value, say, from 10 cm to 200 cm, it becomes definite once the length is set at a particular value. Thus, if the length is set at $l = 90 \text{ cm}$, any measurement of length under the same conditions will yield the same value, regardless of how many times the measurement is repeated (disregarding the small *errors* that may come up in the measurements). Or, think of a small computer program that computes the value of the variable $y = x^2 + 3$, with the value of x fed into the program as an input. Setting the input at $x = 1$, one gets the value $y = 4$, no matter how many times the program is run, though y may have other possible values depending on the value assigned to x . In these examples, l and y are deterministic variables.

On the other hand, if a die with six faces numbered from 1 to 6 is thrown time and again under similar conditions, the number coming up (the ‘reading’ of the die) will keep on changing, and constitutes an instance of a random variable. Here, the reading, say 2, noted in any one throw does not necessarily come up in the next throw where the reading can again be any number from 1 to 6.

Considering a random variable, say, x , the *possible values* of the variable may either be *continuously* distributed over a certain range, or may form a *discrete* set of values. For the sake of simplicity, let us confine ourselves to a consideration of discrete random

variables alone.

Let the possible values of the random variable x be x_1, x_2, \dots, x_n . Suppose that the value of the variable is determined N times, where N is a large number. For instance, a die may be thrown ($N =$) 1000 times, the possible values turning up at any throw being $1, 2, \dots, 6$.

Let, in the N determinations of the value of the random variable x , the value x_i be obtained N_i number of times ($i = 1, \dots, n$). Then, assuming that N is sufficiently large, the ratio $\frac{N_i}{N}$, i.e., the relative frequency of the value x_i being obtained in the N number of trials, is termed the *probability* of the value x_i ($i = 1, \dots, n$). Denoting the probability by p_i , one has the definition

$$p_i = \lim_{N \rightarrow \infty} \frac{N_i}{N} \quad (i = 1, 2, \dots, N). \quad (8-117a)$$

Evidently, the probabilities p_i are non-negative numbers ($p_i \geq 0$) satisfying, among themselves, the relation

$$\sum_{i=1}^n p_i = 1, \quad (8-117b)$$

(reason this out; for sufficiently large N , $\sum_i p_i = \sum_i \frac{N_i}{N}$, where $\sum_i N_i = N$).

An assignment of probabilities for all the possible values of a random variable is referred to as a *probability distribution*. A random variable is characterized completely by the probability distribution over its set of possible values. The probabilities p_i ($i = 1, \dots, n$) making up a probability distribution have to satisfy $0 \leq p_i \leq 1$ ($i = 1, \dots, n$) as also the relation (8-117b), the latter being referred to as the *normalization condition* to be satisfied by the probability distribution.

For instance, for an unloaded die, where all its six faces are equally likely to come up in a throw, the possible values of the number coming up (the ‘reading’ of the die) are $x_1 = 1, x_2 = 2, \dots, x_6 = 6$, and the probabilities for these values coming up are all equal, i.e., the probability distribution is given by $p_1 = p_2 = \dots p_6 = \frac{1}{6}$.

Suppose now that the values taken up by a random variable x are noted in N number of trials, among which N_i trials give the value x_i ($i = 1, \dots, n$). The sum of all these values determined in the N trials is then $\sum_i N_i x_i$. Accordingly, the *average* value of the random variable obtained from the trials will be $\frac{1}{N} \sum_i N_i x_i$. Assuming that N is sufficiently large, and making use of the probabilities p_i characterizing the random variable x , one obtains the following formula for the average value of x determined in a large number of trials

$$\bar{x} = \sum_{i=1}^N p_i x_i. \quad (8-118)$$

This is sometimes referred to as the *expectation value* of the random variable x .

Given the random variable x with possible values x_i and probabilities p_i ($i = 1, 2, \dots, n$), one can determine the expectation value of not only x , but of any *function* of x as well. For instance, x^2 may also be looked upon as a random variable with possible values x_i^2 ($i = 1, \dots, n$), and with the same set of probabilities p_i ($i = 1, 2, \dots, n$) for these possible values. For instance, one may be interested in the squared value of the number turning up in the throw of a die, which can be any one among the numbers 1, 4, 9, 16, 25, and 36. Evidently, the probabilities of these possible squared values for an unloaded die are all $\frac{1}{6}$.

One can now consider the expectation value of this new random variable (call it x^2), the squared value of the random variable x . This is given by the formula

$$\overline{x^2} = \sum_{i=1}^n p_i x_i^2. \quad (8-119)$$

As an example, the instantaneous speed of the molecules of a gas can be looked upon as a random variable, having possible values ranging from 0 to ∞ , the probability density for the speed c being, say $p(c)$.

Since the possible values of the speed are distributed *continuously*, the probability distribution is described by a *probability density* function rather than a discrete set of probabilities. It is defined in such a way that the probability of the speed having any

value within a small interval c to $c + \delta c$ is $p(c)\delta c$. The function $p(c)$ for the speed of a gas molecule at temperature T can be obtained from Maxwell's velocity distribution formula, and has been depicted graphically in fig. 8-12.

The squared speed c^2 is then a random variable as well, with a probability distribution determined by $p(c)$. The mean of the squared speed ($C^2 = \overline{c^2}$) is then an important quantity relating to the state of the gas as a whole, and its square root is referred to as the root mean squared (RMS) speed.

Two distinct probability distributions for a random variable x may give the same expectation value \bar{x} , but they will not necessarily correspond to the same value of $\overline{x^2}$. Thus, $\overline{x^2}$ often serves the purpose of conveniently distinguishing between more than one alternative probability distributions of a random variable.

However, it is still possible that two distinct probability distributions are characterized by the same value of $\overline{x^2}$ as well as of \bar{x} . These can then be distinguished in terms of $\overline{x^3}$, $\overline{x^4}, \dots$, termed the *moments* of various orders of the probability distributions.

Chapter 9

Wave motion I: Acoustic waves

9.1 Simple harmonic oscillations of physical quantities

The present section recalls some of the ideas we met with in section 4.3.

Suppose that the value of a physical quantity varies with time so as to increase and decrease alternately on either side of some mean value. Then such a variation is said to constitute an *oscillation*, or pulsation of that physical quantity. For instance, when a particle undergoes simple harmonic motion, its instantaneous displacement from the mean position alternately assumes positive and negative values, and this variation with time is an example of an oscillation. In this instance, the variation of the velocity of the particle can also be cited as an instance of an oscillation. Pulsations relating to the periodic motion of a particle or a body are sometimes referred to as vibration while the term oscillation is also used in the specific context of such a motion.

As another example, imagine a situation where the pressure in some small region in the atmosphere is varying with time, increasing and decreasing alternately. One can then say that there is taking place an oscillation in the physical quantity termed pressure. In a similar manner, one can talk of oscillations in the temperature of the atmosphere at a particular place, or of oscillations in the current flowing through an electrical circuit.

It is to be mentioned here that oscillations in the value of a physical quantity need not always imply a *periodic* variation, i.e., one in which there occurs a continual repetition of the same set of values of that quantity. As an example, think of the damped simple harmonic motion of a particle in which the amplitude of oscillation decreases gradually with time. Here the state of motion of the particle cannot be precisely said to repeat itself at fixed intervals of time, but still one can say from a broad point of view that the displacement or the velocity of the particle undergoes an oscillation. Indeed, one commonly encounters situations in the physical sciences where measurable quantities of various descriptions undergo oscillatory variations that may not be periodic in a strict sense. We will, however, confine ourselves to the consideration of periodic variations alone for the time being.

The simplest example of periodic oscillations is, of course, a *simple harmonic oscillation*, referred to as simple harmonic motion (SHM in brief) in the context of oscillations of a particle or a body (see chapter 4). Denoting the instantaneous value of the physical quantity under consideration by u , its variation with time (t) will in this case be given by an expression of the form

$$u(t) = a \cos(\omega t + \delta). \quad (9-1)$$

In this expression a stands for the amplitude of oscillation of the physical quantity, ω for the angular frequency (related to the time period (T) of oscillation as $\omega = \frac{2\pi}{T}$), and δ for the initial phase, while $\phi = \omega t + \delta$ represents the phase angle of the oscillation.

9.2 Oscillations transmitted through space: waves

Imagine a physical quantity having some definite value at every point in some region of space, where the value at any given point may possibly change with time as well. As an example, think of the atmospheric pressure at various different points in some region of space. If barometers (instruments for measuring atmospheric pressure) be set up at a number of points in this region and readings be taken of these barometers at any given time instant, it will be found that the readings differ, perhaps by small amounts, from

one another. This means that the pressure at any given point in the region at a given instant of time depends on the location of that point. Again, if one reads the barometer at any fixed location at a number of different time instants, it will be found that these readings are also not all of the same value. This means that the atmospheric pressure is a physical quantity whose value depends not only on location in space but also on time.

Such a situation, where the value of a physical quantity depends on the location in space, is often described by saying that a *field* of the physical quantity under consideration has been set up in the given region of space. Thus, in the above example, one can speak of a *pressure field* having been set up in which, moreover, the pressure undergoes a variation with time as well.

Suppose now that such a field has been set up for a given physical quantity in a given region of space and that, at some given point in this region, an *oscillation* is taking place due to variations in the value of this physical quantity with time. Imagine that, with the passage of time, the oscillation is *transmitted* to neighboring points of space in that region and that, in this manner, the oscillation spreads to more and more distant points. Such a situation is described by saying that a *wave* of the physical quantity under consideration has been set up in that region of space. In other words, a wave can be defined as *a transmission of an oscillation from one location to various other locations in space*.

As I have already indicated, oscillations can be of various different types, corresponding to physical quantities of different descriptions. Moreover, the manner in which the oscillations are transmitted through space can also differ in different instances. As a result, denoting the value of the physical quantity under consideration by u , the nature of space- and time dependence of u may be of various different types, corresponding to waves of different kinds. The space- and time dependence is usually denoted by expressing u as a function $u(\mathbf{r}, t)$, where t denotes time, and \mathbf{r} denotes the position vector of any given point in space. Denoting the components of \mathbf{r} with respect to any given Cartesian co-ordinate system by x , y , and z , one can represent the above functional

dependence alternatively as $u(x, y, z, t)$. The nature of this functional dependence differs for waves of different descriptions. The function u is sometimes referred to as the *wave function* characterizing the wave under consideration.

Though I will occasionally use the term wave function in the above sense, this term is more commonly used in the specific context of quantum theory, where it is employed to denote a function characterizing the instantaneous state of a microscopic system (see chapter 16).

With this introduction to the concept of waves we will, in this book, look at examples of waves in a number of specific areas in physics in greater detail, indicating a few of the more important features of these waves. Our first topic, to be covered in the present chapter, will be *acoustic* or sound waves where, we will also refer to *elastic waves* for the sake of generality since, to be precise, sound waves are a special case of elastic waves. Next, in chapter 14, we will look at *electromagnetic waves*, of which *light waves* (chapter 15) constitute a special case.

The term ‘acoustic’ is generally used to refer to audible sound. Elastic pressure waves (see sec. 9.3) with frequencies ranging from 30 Hz to 30 kHz (both approximate figures) and with small amplitudes are usually referred to as acoustic waves, though the range of audible frequencies varies from person to person.

9.3 Sound waves as variations in pressure: elastic waves

When we hear a sound, some vibrating body or other acts as the source of sound and part of its energy of vibration is transmitted through some medium to our ears in the form of a wave, generating the sensation of sound, where the ear acts as a receiver or sensor of sound. The medium in which the waves are set up will, for the present, be assumed to be air. The transmission of sound waves in other media will be referred to later.

The vibration of the source of sound disturbs the equilibrium of the layer of air adjacent

to the source. Under the impact of the periodic motion of the source, the pressure in this adjacent region undergoes a periodic rise and fall within certain small limits. In turn, this variation of pressure causes a similar periodic variation in the pressure of the layer of air next to it. In this manner, periodic pressure variations are transmitted to distant regions through the 'impact' of each layer on contiguous layers. As I mentioned earlier, a wave means the transmission of an oscillation of some physical quantity across some region of space. In this present instance, the transmission of sound is nothing but the setting up of a *pressure wave* in air.

As we know from chapters 6 and 7, the elastic property of air or any other fluid is characterized by the special feature that whatever be the strain produced in it, the stress is always in the nature of a volume or bulk stress, and the value of this volume stress is simply the pressure with a negative sign ($-p$). Hence the transmission of sound waves in air can be described in a slightly different language: due the impact of the vibrating source, a strain is set up in the layer of air adjacent to it and, as a response, a bulk stress is generated in this layer. In turn, the impact of this layer on the next adjacent one (or, more precisely, the momentum transfer between the two layers), causes a strain and a resulting bulk stress to be set up in the latter. It is in this manner that an *elastic wave* spreads out from the source to distant regions. This is nothing but the pressure wave I mentioned above. If the frequency of the wave lies in a certain range, we perceive it as sound. Additionally, the waves are required to be of sufficiently small amplitude, which allows one to describe these in simple terms. Waves of larger amplitude have a number of features not commonly observed in ordinary sound, and require more complex modes of description.

Elastic waves can also be generated in media other than air due to the impact of vibrating sources. If the medium be made up of a fluid, say, water or some other liquid, then the elastic wave happens to be of a rather simple nature, namely a pressure wave. However, in solid media, the spreading of elastic waves is a more complex process. I will briefly refer to this in a later section.

In order to explain the transmission of sound in the form of a pressure wave, let us

refer to a vibrating tuning fork and look at fig. 9-1 (A), (B) below. In fig. 9-1(A), alternate regions of high and low pressure in successive layers of air near one arm of the tuning fork have been shown schematically at some particular instant of time, high pressure regions being indicated by the letter 'h' and regions of comparatively low pressure by 'l'. On the other hand, in 9-1 (B), the same layers have been shown after a short lapse of time. One observes that high pressure layers have changed to low pressure ones while low pressure regions have been converted into high pressure ones, where the words 'high' and 'low' have been used in a relative sense, to denote changes in pressure within some small range.

One can describe this situation by saying that the *phases* of high and low pressure are gradually transmitted from the tuning fork away to more and more distant regions, the term 'phase' being used here to describe the instantaneous state of oscillation of a physical quantity. Thus, for instance, the low pressure phase in the layer A of fig. 9-1(A) has shifted to layer A' in 9-1(B) while the high pressure phase at B has shifted to B'. Looking at any particular layer, say, A, at a fixed location in space, it will be seen that the changing phase at the location of the layer results in an alternate increase and decrease of pressure in it.

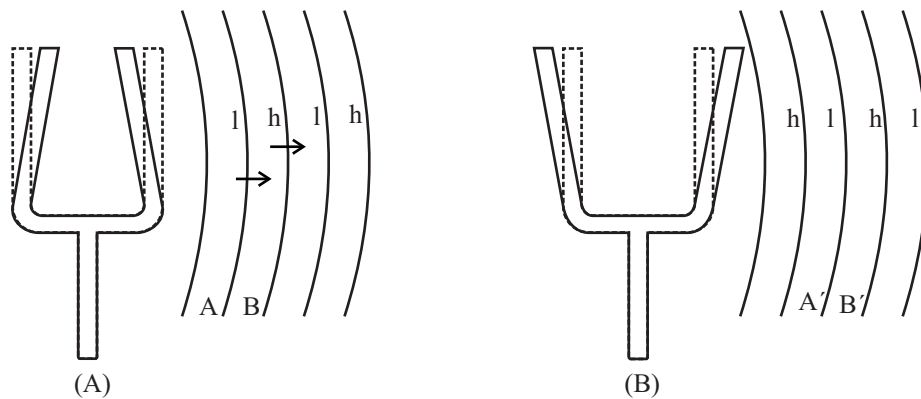


Figure 9-1: Layers of high and low pressure adjacent to one arm of a vibrating tuning fork (schematic); the low- and high pressure layers A and B in (A) have shifted, after a small time interval, to A' and B' in (B) in the direction shown by arrows; such layers are denoted by symbols 'l' and 'h' respectively, which alternate in space at any given instant of time.

9.4 Sound waves in one dimension

9.4.1 Variation of excess pressure

The sound from a tuning fork actually spreads in all directions around it, though this has not been shown in fig. 9-1(A), (B). The mathematical description of such a wave is somewhat involved. Instead, we will think of a wave that spreads only along one direction, say, along the x -axis. In that case, the pressure (P) at any point on the x -axis will depend on the co-ordinate (x) of that point and also on the time (t). In this context, it is convenient to use, instead of the pressure P , the *difference* (p) between this pressure and the *mean pressure* P_0

The mean pressure P_0 is simply the air pressure in absence of the sound wave, while p represents the *excess pressure* arising due to the wave being set up. This excess pressure (also referred to as the *acoustic pressure*) goes through positive and negative values due to the instantaneous pressure P alternately rising above and falling below the mean pressure P_0 .

The pressure at any point and any given instant of time is the sum of the mean pressure and the excess pressure. The following picture is to be kept in mind while describing and analyzing sound waves set up in a medium.

We consider a small volume element within the fluid medium in which the acoustic wave is set up where, for the sake of concreteness, the medium is assumed to be air. Though of a small size, the volume element is still assumed to be of macroscopic proportions so as to contain a large number of microscopic constituents (molecules) of the medium. In other words, we make the assumption that there exist infinitesimally small volume elements in the fluid made up of large numbers (of macroscopic magnitude) of molecules, where such an assumption is found to lead to meaningful results in spite of its idealized nature. At times, such a small volume element is referred to as a fluid ‘particle’ (refer to discussion in section 7.3.1).

Considering a such small volume element around any given point in the medium, the molecules in this small element can be assumed to be in a state of thermal equilib-

rium *among themselves*, though the element as a whole may *not* be in a state of mechanical and thermal equilibrium with *surrounding* volume elements. In other words, each small volume element may be assumed to be characterized by its own equilibrium thermodynamic parameters (such as pressure, density and temperature) which, however vary from one element to another because of a non-equilibrium situation prevailing among these various volume elements because of the acoustic wave set up in the medium. One can look upon each such volume element as a macroscopic system, instantaneously in equilibrium, where the state of equilibrium undergoes a periodic variation due to interactions with other surrounding elements of a similar description. As regards the microscopic constituents making up the volume element, there takes place a constant influx and out-flux of individual molecules through its boundary surface, though its macroscopic identity remains unchanged since the mean number of constituents (or, more precisely, the *mole* number) remains unchanged.

As an acoustic wave passes through the medium under consideration, each volume element executes mechanical oscillations about its mean position and, at the same time, suffers alternate compression and dilatation, whereby the elastic stress within the element (in the form of pressure) varies periodically.

The excess pressure (p), which is the wave function in this case, is a function of position (x) and time (t). If the vibration of the source and the resulting variation of the excess pressure is characterized by one single frequency then the expression for p will be of the following form:

$$p = p_0 \cos \left(2\pi \left(\frac{x}{\lambda} - \frac{t}{T} \right) \right). \quad (9-2)$$

In this expression p_0 indicates the amplitude of the excess pressure, i.e., the maximum magnitude of the difference between the instantaneous pressure (P) and the mean pressure (P_0), while T stands for the time period of the vibration of the source and of the resulting sound, and λ for the *wavelength* (see below) of the acoustic wave.

The wavelengths of acoustic waves set up in air typically lie in the range $\sim 10^{-2}$ m to ~ 10 m.

If one follows the rise and fall of pressure (or of the excess pressure) at any given point, i.e., for any particular value of x , then the value of pressure will be found to repeat itself at some definite time interval, which is then identified as the time period. On the other hand, the x -interval at which the value of the pressure is repeated at any given time instant, is the wavelength. Figure out how these two definitions check with equation (9-2).

The symbols ω and k are often used for $\frac{2\pi}{T}$ and $\frac{2\pi}{\lambda}$, and are termed, respectively, the angular frequency and the wave number:

$$\omega = \frac{2\pi}{T}, \quad k = \frac{2\pi}{\lambda}. \quad (9-3)$$

In terms of these two quantities the expression (9-2) for the excess pressure reads

$$p = p_0 \cos(kx - \omega t). \quad (9-4)$$

The expression $(kx - \omega t)$ is referred to as the *phase angle* or simply *phase* at the point x at time t . Denoting the phase by $\Phi(x, t)$, (9-4) can be written as

$$p = p_0 \cos \Phi(x, t). \quad (9-5)$$

Note that the wave function, i.e., the excess pressure, depends on the phase through its trigonometric cosine, as a result of which the value of the excess pressure repeats itself as the phase changes by 2π . In other words, one can ignore integral multiples of 2π occurring additively in the value of phase.

For instance, replacing t with $t + T$, or x with $x + \lambda$, results in a change of 2π in the value of the phase, which explains the above definitions of time period and wavelength.

While writing (9-2) or (9-4) we have assumed that the phase is zero at the point $x = 0$ at the instant $t = 0$. However, this is not essential and one can as well add a constant, say

δ , to Φ to obtain a more general form for the excess pressure:

$$p = p_0 \cos(kx - \omega t + \delta). \quad (9-6)$$

This does not change the basic fact that the value of the excess pressure repeats itself at intervals of T in time and λ in position. While (9-4) is nothing but (9-6) with $\delta = 0$, the choice $\delta = -\frac{\pi}{2}$ leads to

$$p = p_0 \sin(kx - \omega t), \quad (9-7)$$

and $\delta = \frac{\pi}{2}$ to

$$p = p_0 \sin(\omega t - kx). \quad (9-8)$$

In other words, the expression for the wave function (excess pressure in this case) can be given alternative forms with various different choices of δ , referred to as *initial phase* or *epoch*.

9.4.2 Propagation of the monochromatic wave

Figure 9-2 depicts graphically the variation of excess pressure with time (t) at any given point in space, say, at $x = 0$, in accordance with eq. (9-4). Evidently, it represents a periodic variation with period T indicated in the figure since, for instance, the variation of p during the interval $t = 0$ to $t = T$ is exactly similar to that during $t = T$ to $t = 2T$. Indeed, the graph in fig. 9-2 is similar to that in fig. 4-3(A) because the variation of p with time is a simple harmonic oscillation.

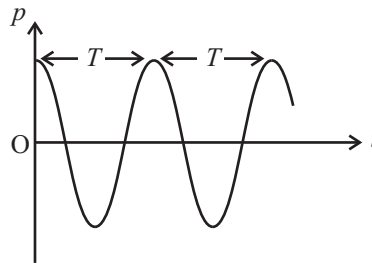


Figure 9-2: Graph depicting variation of excess pressure with time at any given point in space; the variation is periodic, with period T .

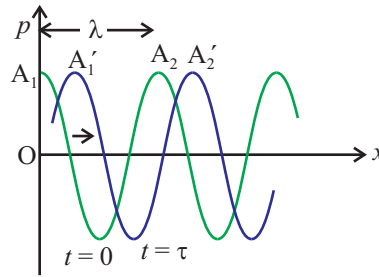


Figure 9-3: Graph depicting variation of excess pressure with position x for two instants of time, $t = 0$ and $t = \tau$; the variation is periodic with period λ , the wavelength; in the lapse of time from $t = 0$ to $t = \tau$, the waveform gets shifted through a distance $\frac{\lambda\tau}{T}$.

Now look at fig. 9-3. This figure depicts graphically the variation of excess pressure with the spatial co-ordinate x at a given time instant, say $t = 0$ and also at a later time, say $t = \tau$. Notice that the variation in this figure is again a sinusoidal one, being of the same nature as in fig. 9-2. Here the value of p is repeated at an x -interval λ - the wavelength of the wave under consideration. The points on the graph corresponding to the maximum possible value of p (in this instance, this maximum value is p_0) at the given instant of time ($t = 0$) are termed 'crests'. These correspond to points like A_1 and A_2 in the figure, and the distance between successive crests is seen to be λ (check this out from equation (9-4) or, more generally, from (9-6)). Similarly, the points corresponding to minimum possible values ($-p_0$ in this instance) of p are termed 'troughs' - successive troughs are once again separated by the wavelength λ .

Notice from the figure that the graph for $t = \tau$ can be obtained by simply shifting the graph for $t = 0$ through a certain distance along the x -axis. In this process, the crests of the first graph all coincide with those ($A'_1, A'_2 \dots$) of the second graph and similarly, there occurs a coincidence of the set of troughs as well. The graph depicting the variation of the wave function with spatial distance (the co-ordinate x in the present instance) at some particular time instant is sometimes referred to as the *waveform*, or *wave profile*. Thus, one can say that in the interval from $t = 0$ to $t = \tau$ the entire waveform gets shifted through a certain distance. If, in particular, τ equals the time period T , then the waveform gets shifted by λ and the crests A_1, A_2 , etc. coincide respectively with the succeeding crests of the *same* waveform (A_1 with A_2 , etc.). Thus, the *rate* of displacement

of the waveform with time is $\frac{\lambda}{T}$. This is termed the *phase velocity* of the wave (the term *wave velocity* is also used). This means that in time τ the waveform gets displaced through the distance $\frac{\lambda\tau}{T}$.

Recall that the value of the wave function (the excess pressure p in the present instance) at any point in space at any given instant of time is determined by the *phase*. If the phase is Φ at the point x at $t = 0$, then the same value of the phase will occur at the point $x + \frac{\lambda\tau}{T}$ at time $t = \tau$ (check this out). The points in space at which the phase equals some given value, say Φ , at any given time instant are said to constitute a *wave front* at that instant. Different possible values of the phase then correspond to various different wave fronts at that instant. In the present instance, the wave fronts are nothing but points on the x-axis (or, planes perpendicular to the x-axis, if the wave is considered as being set up in three dimensional space; see sec. 9.5). The shift or displacement of the waveform with the passage of time can then alternatively be described as a translation of the wave fronts. Thus, in fig. 9-3, the value of the phase at the point A_1 ($x = 0$), which is $\Phi = 0$ according to equations (9-4), (9-5), recurs at the point A'_1 at $t = \tau$, i.e., the wave front located at A_1 has shifted to A'_1 in time τ . The phase velocity is then simply the *rate of displacement of the wave fronts*.

The above expression for the phase velocity of the wave can be given an alternative form in terms of the angular frequency ω and wavenumber k instead of time period T and wavelength λ by making use of the relations (9-3), while the relation $T = \frac{1}{\nu}$ (where ν stands for the frequency) can also be used to express the same in terms of ν and λ :

$$v = \frac{\lambda}{T} = \frac{\omega}{k} = \nu\lambda. \quad (9-9)$$

In the above paragraphs we have considered the particular instance of a wave where the physical quantity represented by the wave function is the excess pressure because it is the excess pressure that undergoes oscillatory variations at various different points in space as an acoustic wave is set up. As I have mentioned above, a wave can also be set up when some *other* physical quantity that undergo is made to oscillate in some region of space (in particular, in the case of an acoustic wave, there

exist *other* physical quantities as well that undergo oscillatory variations along with the pressure). The wave that we have been considering here is of a simple type, where the wave function depends only on x and t (there being no dependence on the other two spatial co-ordinates y and z), and the variation of the wave function is sinusoidal (equations (9-2), (9-4)).

Regardless of whether the wave function is the excess pressure or some other physical quantity, however, the mathematical description and characteristics of the wave will remain the same, while the physical interpretation of the mathematical equations will differ from one context to another. Waves of more complex nature than that expressed by (9-2), (9-4) may also propagate in space, where the expressions describing the variations of the wave functions look more involved.

The wave expressed by eq. (9-6) propagates in the *positive* direction of the x -axis. A wave can equally well propagate along the *negative* direction of the x -axis, in which case its wave function would have been given, instead of (9-6), by

$$p = p_0 \cos(kx + \omega t + \delta). \quad (9-10)$$

This is explained as follows. The value of the phase at the point $x = 0$ and at time $t = 0$ is $\Phi = \delta$. After an interval τ (with, say, $\tau > 0$), the same value of the phase occurs at the point $x = -\frac{\omega\tau}{k}$, as can be seen directly from (9-10) by substitution. This means that in time τ the wave front corresponding to $\Phi = \delta$ gets shifted towards the negative direction of the x -axis through a distance $\frac{\omega\tau}{k}$. In other words, the rate of displacement of the wave front is $\frac{\omega}{k}$ along the direction of x decreasing. The same conclusion can be seen to follow for any other value of the phase Φ .

The waves described by the formula (9-6) or (9-10) is referred to as a plane progressive monochromatic wave. The term ‘monochromatic’ signifies that the time variation of the wave function occurs with a single frequency ($\nu = \frac{\omega}{2\pi}$), while the term ‘progressive’ indicates that the wave profile propagates with a certain specific velocity. The term

‘plane’ indicates that, considered as a wave in three dimensions, the *wave fronts* are plane surfaces. I will explain this more fully in sec. 9.5.

9.5 Waves in three dimensions

9.5.1 The plane progressive wave

A wave propagating along the positive or negative direction of the x-axis can be termed a *one dimensional* wave. However, instead of the x-axis, a one dimensional wave can propagate in some other direction of space as well. Consider, for instance, a unit vector \hat{n} in some arbitrarily chosen direction in space. Then a sinusoidal wave propagating along this direction will be of the form

$$p = p_0 \cos(\mathbf{k} \cdot \mathbf{r} - \omega t + \delta). \quad (9-11)$$

In this expression, the vector \mathbf{k} stands for $k\hat{n}$, and is referred to as the *wave vector*.

1. For a wave corresponding to some physical quantity other than the excess pressure, one has to substitute appropriate symbols in place of p and p_0 .
2. If the direction of propagation happens to be along the positive or negative direction of the x-axis, one has to take $\hat{n} = \hat{i}$ or $\hat{n} = -\hat{i}$, and then the expression (9-11) reduces to (9-6) or (9-10) respectively.

Notice that the wave function in (9-11), as also the phase, depends, in general, on three space variables (x, y, z) instead of just one, in addition to the time variable t :

$$\Phi = k\hat{n} \cdot \mathbf{r} - \omega t + \delta. \quad (9-12)$$

Referring to any particular value of Φ and any particular instant of time t , the points in space where the phase happens to have that particular value, can be seen to lie on a *plane*. As I have mentioned above, a set of such points are said to constitute a *wave front* at that instant of time. In other words, a typical wave front for the wave described by (9-11) is a plane. This is why the wave described by (9-11) is termed a *plane wave*.

Further, all the plane wave fronts corresponding to various different values of Φ can be seen to be perpendicular to the unit vector \hat{n} and hence, all these wave fronts are parallel to one another (see fig. 9-4). The unit vector \hat{n} is referred to as the *wave normal*. Comparing the positions of the wave fronts at different instants of time, it can be seen that each wave front is translated in space in a direction parallel to \hat{n} with a phase velocity $v = \frac{\omega}{k}$, describing the transmission of the oscillations of the wave function in space.

Though a wave represented by eq. (9-6) or (9-10) is termed a one dimensional wave, it may in reality correspond to a wave propagating in three dimensions where, as a special case, the co-ordinates y and z , are absent in the expression for Φ . The wave fronts in this special case are planes perpendicular to the x-axis, propagating along the positive or the negative direction of the x-axis with the phase velocity of the wave.

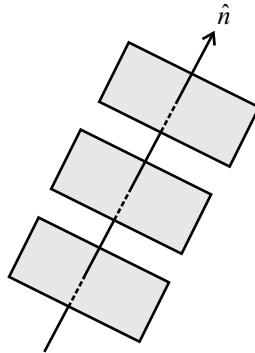


Figure 9-4: A set of plane wave fronts parallel to one another; \hat{n} is perpendicular to all the wave fronts, and gives the direction of displacement of the latter as these move with the passage of time.

In a more complete description the wave (9-11) is termed a *plane monochromatic traveling wave*, while the terms ‘propagating’ or ‘progressive’ are also used in place of the term ‘traveling’. Here the term *monochromatic* refers to the fact that the oscillations of the wave function occur with one single frequency ν (angular frequency ω) as a result of which the waveform describing the oscillations is sinusoidal. In contrast, a wave that can be described as a superposition of a number of monochromatic waves (see below) is

termed a polychromatic (or, non-monochromatic) wave.

The origin of the terms monochromatic and polychromatic lies in optics. The sensation of *colour* generated by a light wave depends on the frequency (or the set of frequencies) characterizing the wave.

Problem 9-1

Show that the expression

$$P = A + B \cos^2(ax + by - ct),$$

where P stands for the pressure and A, B, a, b, c are constants, represents a plane monochromatic wave, and find the mean pressure, amplitude, propagation direction, frequency, and phase velocity of the wave.

Answer to Problem 9-1

HINT: One can write the given relation in the form

$$p \equiv P - \left(A + \frac{B}{2}\right) = \frac{B}{2} \cos(2ax + 2by - 2ct).$$

Comparing with eq. (9-11), where p stands for the excess pressure, one finds that the above expression indeed represents a plane monochromatic wave where the mean pressure is $A + \frac{B}{2}$, the amplitude is $\frac{B}{2}$, and the propagation vector $\mathbf{k} = 2a\hat{i} + 2b\hat{j}$ has direction cosines $(\frac{a}{\sqrt{a^2+b^2}}, \frac{b}{\sqrt{a^2+b^2}}, 0)$. The angular frequency of the wave being $2c$, its frequency is $\frac{c}{\pi}$, and the phase velocity of the wave is $v_p = \frac{\omega}{|\mathbf{k}|} = \frac{c}{\sqrt{a^2+b^2}}$.

9.5.2 Waves of more general types

As I have mentioned, there may be waves of various different types *other* than the plane monochromatic wave introduced above. Depending on the context, the expression for the wave function describing its variation in time and space may be more or less complex. From a mathematical point of view one can say that a wave is set up when some physical quantity (say, u ; in the case of a sound wave, this is the excess pressure p)

varies in space and time in accordance with a *partial differential equation* belonging to a certain category. I will not enter here into a general discussion of the partial differential equations describing waves of various types but I must mention that sound waves, which constitute a particular instance of elastic waves of small amplitude, and light waves (or, more generally *electromagnetic waves*, see chapter 14) can be described from this more general mathematical viewpoint.

9.5.3 The principle of superposition

The plane monochromatic traveling wave is just a specific type of solution of such a partial differential equation. The partial differential equations describing such waves possess the special feature that if $u_1(x, y, z, t)$ and $u_2(x, y, z, t)$ are any two solutions of the equation describing two waves set up in space, then $u_1(x, y, z, t) + u_2(x, y, z, t)$ is also a wave solution satisfying the same equation. This is termed the *principle of superposition of waves*, and one then says that the wave described by $u_1 + u_2$ is formed by the superposition of the waves u_1 and u_2 .

The partial differential equation describing acoustic waves constitutes an instance of a class of equations generically referred to as the wave equation. This we will take up in sec. 9.6.

Let me illustrate the principle of superposition here with an example. Let A and B denote two independent sources of sound. Suppose that $p_1(\mathbf{r}, t)$ represents the excess pressure due to the wave generated by A in the absence of B while, similarly $p_2(\mathbf{r}, t)$ stands for the excess pressure due to the wave generated by B in the absence of A. Then the excess pressure corresponding to the wave set up by A and B vibrating together would be given by $p_1(\mathbf{r}, t) + p_2(\mathbf{r}, t)$. However, if the amplitude or intensity of the sound be of a large magnitude then the principle of superposition may not be applicable (on the other hand, this principle is always applicable for electromagnetic waves set up in vacuum regardless of the amplitude and intensity). The failure of the principle of superposition in the case of waves of large amplitude is responsible for a number of novel effects such

as the production of waves whose frequencies are *combinations* of the frequencies of two sources of different frequencies when these act one in the presence of the other (an effect which is to be distinguished from the production of beats (refer to sec. 9.15.3), which is an effect resulting from the superposition of waves of small amplitudes.).

The principle of superposition allows us to construct expressions for waves of various different descriptions by making up superpositions of more than one plane progressive monochromatic waves. These waves with more or less complex expressions for their wave functions can then be made use of in describing and explaining phenomena relating to wave propagation in various physical situations of interest.

In talking of the principle of superposition, it is necessary to make a mention of *boundary conditions*. For instance, suppose that the function u_1 satisfies the wave equation (like the one describing the propagation of sound, see sec. 9.6) within the volume of a sphere, and satisfies the boundary condition of being zero on the surface of the sphere (the boundary condition expresses some constraint on the wave function depending on the physical situation under consideration; see sec. 9.6 and 9.7). Suppose also that the function u_2 satisfies the same equation within the volume of the sphere, and satisfies the *same* boundary condition. Then $u_1 + u_2$ will also be a solution, satisfying, once again, the same boundary condition. If, however, u_1 and u_2 satisfy a pair of two different boundary conditions, then the superposition $u_1 + u_2$ will satisfy neither of these.

9.6 The wave equation

Consider a one dimensional wave represented by the wave function $u(x, t)$ where, analogous to eq. (9-6)

$$u = u_0 \cos(kx - \omega t + \delta), \quad (9-13)$$

and where u_0 is the amplitude of the wave. Here, the function $u(x, t)$ satisfies the *differ-*

ential equation

$$\frac{\partial^2 u}{\partial x^2} - \frac{1}{v^2} \frac{\partial^2 u}{\partial t^2} = 0, \quad (9-14)$$

as can be verified by direct substitution, with $v = \frac{\omega}{k}$, the phase velocity of the wave. In this equation, the second order *partial* derivatives represented by $\frac{\partial^2}{\partial x^2}$ and $\frac{\partial^2}{\partial t^2}$ occur, which are obtained by successive applications of the first order partial derivatives $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial t}$ respectively. A partial derivative such as $\frac{\partial}{\partial x}$ corresponds to taking the derivative with respect to the variable x while the other relevant variable t is held fixed at any given value.

1. Thus, for instance,

$$\frac{\partial}{\partial x} \cos(kx - \omega t + \delta) = -k \sin(kx - \omega t + \delta), \text{ and } \frac{\partial}{\partial x^2} \cos(kx - \omega t + \delta) = -k^2 \cos(kx - \omega t + \delta).$$

2. An equation such as eq. (9-14) is referred to as a *partial differential equation* because it involves partial derivatives of the unknown function (u in the present instance). The function given by eq. (9-13) then represents a *solution* of the partial differential equation. In general, a partial differential equation may possess numerous different solutions, depending on the *boundary conditions* or *initial conditions* satisfied by these. For instance, a condition of the form

$$u = a \text{ at } x = x_0 \text{ for all } t \text{ (} a, x_0 \text{ given constants)}$$

is termed a boundary condition, while one of the form

$$u = b \text{ at } t = t_0 \text{ for all } x \text{ (} b, t_0 \text{ given constants)}$$

is referred to as an initial condition.

Analogous to eq. (9-14), commonly referred to as the *wave equation in one dimension*, one can also write down the *wave equation in three dimensions*:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} - \frac{1}{v^2} \frac{\partial^2 u}{\partial t^2} = 0. \quad (9-15)$$

In general, the parameter v in eq. (9-14) or (9-15) depends on the medium in which the wave is set up, and possibly on the type of the wave as well. It is given by the formula $v = \frac{\omega}{k}$ for a plane monochromatic wave, and represents its phase velocity.

The wave equation in *two dimensions* will be found in sec. 9.16.2 in the context of vibrations of a stretched diaphragm.

As I have already mentioned, the plane monochromatic wave constitutes just one particular solution of the wave equation (9-14) or (9-15), and wave solutions of these equations of a great many other forms are possible. Indeed, one sometimes *defines* a wave as a solution to a wave equation.

The wave equation (9-15) is a *linear* partial differential equation in that it does not involve terms of the second degree or higher of the the wave function u and its derivatives. It is this characteristic of the wave equation due to which the principle of superposition holds. In consequence, *superpositions* of acoustic waves of relatively simple types (such as the plane monochromatic wave or the spherical wave) yield more complex ones.

What is more, wave equations more general than eq. (9-14) or (9-15) are also possible, implying an even broader definition of the concept of waves. In this book, however, we will be concerned with waves of relatively simple description, all satisfying equations of the form (9-14) or (9-15).

9.7 Sources and boundary conditions

The nature of the wave set up in a given real life situation depends on the nature of the *source* of the wave and also on the *boundary conditions* corresponding to it.

9.7.1 The monopole source. Spherical wave.

For instance, if the source is a point-like one and its oscillations occur with one single frequency (e.g. a small balloon undergoing alternate expansion and contraction; such a source is referred to as a *monopole*) in homogeneous and infinitely extended space without the presence of any other intervening body, then it sets up a monochromatic *spherical* traveling wave. Here all the wave fronts happen to be spherical in shape (fig. 9-5) and the radii of the spheres increase at a uniform rate with time. The excess pressure field $p(\mathbf{r}, t)$ is of the form

$$p = \frac{p_0}{r} \cos(kr - \omega t), \quad (9-16)$$

where p_0 is a constant and $k = \frac{2\pi}{\lambda}$, λ being the wavelength measured along the radial direction.

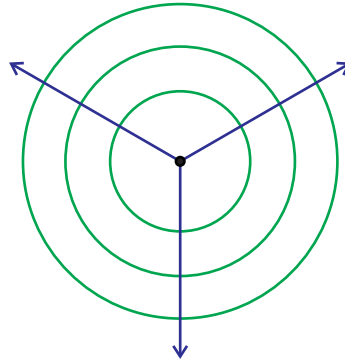


Figure 9-5: Wave fronts for a spherical wave from a monopole source; spheres of successively larger radii correspond either to wave fronts for different phases at the same instant of time or wave front for a given phase at successive instants of time; the arrows indicate wave normals, along which the wave fronts propagate.

The rate of expansion of the spherical wave fronts is $v = \frac{\omega}{k}$ and represents the phase velocity of the spherical wave. As seen from eq. (9-16), the amplitude of the wave varies inversely with r , the distance from the monopole source, corresponding to which the intensity of sound (see sec. 9.11.4 for an introduction to the idea of sound intensity) varies inversely as the *square* of the distance.

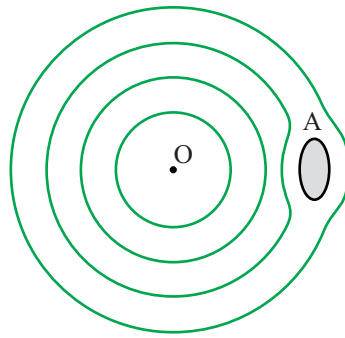


Figure 9-6: Deformation of wave front in the presence of an obstacle; the spherical wave fronts from the monopole source O get deformed in moving around the object A ; the boundary of the object imposes restrictions on the variation of the wave function, leading to the deformation.

In the presence of some other body or obstacle in the vicinity of the source, a number of boundary conditions are imposed on the wave at the boundary surface of that body. As a consequence, the wave breaks up or spreads around that body and the wave fronts assume a different shape (see, e.g., fig. 9-6).

In our everyday experience we perceive sound waves originating from numerous vibrating sources around us. Pressure waves of various descriptions are generated from these sources. If the frequency of any of these waves lies in a certain range (roughly, 30Hz to 30 kHz; the unit of frequency is Hz or s^{-1}) then it produces the sensation of sound in us. Depending on the nature of the wave, the sources of sound can be grouped in any one of various classes. For instance, a monopole source produces a spherical wave as indicated above. A small boxed loudspeaker emitting sound of low frequency can be said, in an approximate sense, to constitute such a monopole source.

Since a small area on the surface of a sphere, with a linear dimension small compared to the radius, can be effectively assumed to be a plane, a spherical wave behaves effectively as a plane wave over limited regions of space at large distances from a monopole source (see fig. 9-7). In a sense, the monochromatic plane wave and the monochromatic spherical wave are the simplest wave patterns consistent with the wave equation (9-15). Sources of other types generate more complex spatial distributions of excess pressure (the time dependence is assumed to be harmonic, with a given frequency ω).

9.7.2 Dipole and quadrupole sources

Analogous to the monopole source, one can talk of *dipole* and *quadrupole* sources, (or sources of higher *multipolarity*, as these are called), where each of these generates a sound pattern (i.e., space- and time dependence of the excess pressure, or the intensity distribution) with specific characteristics of its own. A sound pattern of arbitrary complexity can then be described as a *superposition* of the patterns generated by such sources of various multiplicities.

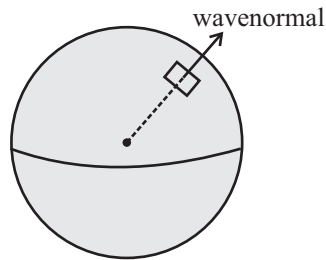


Figure 9-7: A small area on a spherical wave front that looks like the wave front of a plane wave, with the wave normal along the radial direction.

Imagine two small monopole sources located side by side, a small distance apart, with the vibrations of one source differing from that of the other by the phase angle π . In the case of a pair of balloon-like sources referred to above, this means that one of the balloons undergoes a contracting phase when the other is expanding, and *vice versa*. The vibrations of the two sources are then said to occur *in opposite phases*. Such a pair is said to constitute a *dipole* source.

The excess pressure field $p(\mathbf{r}, t)$ for such a source is somewhat different from that of a spherical wave emitted by a monopole source, since the amplitude, in addition to having a dependence on the radial distance from the source, has an angular dependence as well.

A good way to describe the sound from any given source is to make an angular plot, at any given radial distance, of the sound intensity where the latter is usually plotted on a logarithmic scale. The intensity of sound is a quantity directly related to the time-

averaged modulus squared of the excess pressure (see sec. 9.11.4). Fig. 9-8 depicts such a plot for a dipole source, in which one finds that the sound from a dipole source is emitted preferentially along the dipole axis (the line XOX' joining the two oscillating monopoles) while it falls off away from the axis.

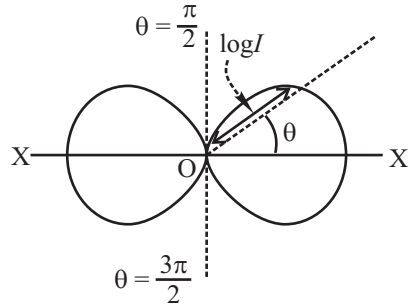


Figure 9-8: Angular plot of the intensity (I , expressed in a logarithmic scale) distribution for a dipole source; the angle θ is measured from the line XOX' , the dipole axis, and the plot depicts the variation of $\log I$ with θ ; the intensity is maximum in the two directions along the axis ($\theta = 0, \pi$), and a minimum along the two perpendicular directions ($\theta = \frac{\pi}{2}, \frac{3\pi}{2}$).

Instead of a pair of monopoles oscillating harmonically in time, a dipole source can also be made up of a single monopole source oscillating harmonically in time and, at the same time, executing a simple harmonic vibration of a small amplitude along a straight line, the line of vibration constituting the dipole axis.

Now imagine a pair of dipole sources described above, placed a small distance apart. Such a pair is said to constitute a *quadrupole* source. If the axes of the dipole sources are at right angles to one another, one has a *transverse* or lateral quadrupole, while a *longitudinal* or linear quadrupole corresponds to dipole axes pointing along the same line. A vibrating tuning fork is effectively a monochromatic longitudinal quadrupole source (see fig. 9-9).

Fig. 9-10 depicts schematically the angular dependence of the intensity (logarithmic plot) for a linear quadrupole source where the intensity pattern is seen to differ in (A) the *near* and (B) the *far* zones. The near zone is characterized by the condition $kr \ll 1$

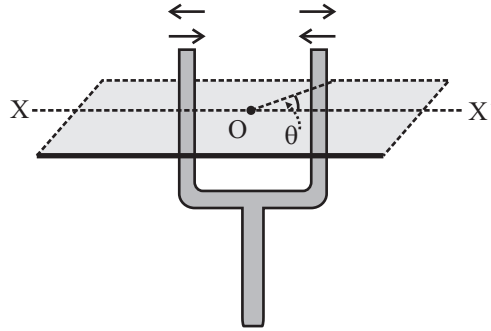


Figure 9-9: The tuning fork as a source of sound; the two tines or arms of the tuning fork vibrate as shown by the arrows, alternately moving towards and away from each other; the vibrating tuning fork effectively acts as a linear quadrupole source; in a plane perpendicular to the lengths of the arms, the line XOX' depicts the axis of the quadrupole; the intensity of sound varies with the angle θ from the axis as in fig. 9-10.

while the far zone corresponds to $kr \gg 1$, where r stands for the radial distance from the source and k is defined as $k = \frac{\omega}{v}$, v being the phase velocity of the acoustic wave emitted by the source.

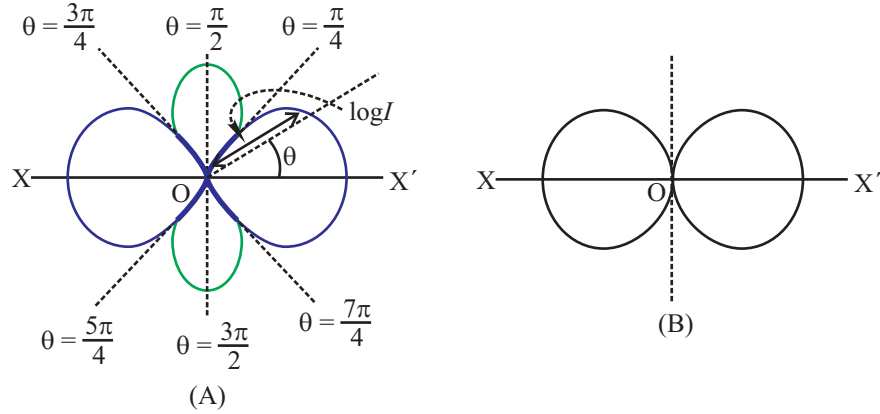


Figure 9-10: Angular distribution of intensity (logarithmic scale) for a linear quadrupole source: (A) near zone and (B) far zone; XOX' denotes the quadrupole axis.

One finds from the figure that, in the near zone, the sound intensity has maxima along the axis of the linear quadrupole (the line joining the two tines in the case of a tuning fork), corresponding to angles $\theta = 0$ and $\theta = \pi$ in the figure, as also along directions perpendicular to the axis ($\theta = \frac{\pi}{2}, \frac{3\pi}{2}$; these two maxima are not as strong as those in the axial direction), and minima in between. In the far zone, on the other hand, there

are maxima along the axial line and minima along the directions perpendicular to it, analogous to the directivity pattern of a dipole source, where the latter resembles a spherical wave with an angular dependence.

Problem 9-2

Propagating acoustic waves are set up in a medium with expressions for excess pressure given by $p_1(x, y, z, t) = A \cos(kx - \omega t)$, $p_2(x, y, z, t) = A \cos(ky - \omega t)$, where x, y, z refer to the co-ordinates in a Cartesian system. Find the expression for excess pressure of the wave resulting from the superposition of the two waves, and show that it is a plane wave (refer to eq. (9-11)) propagating along the direction bisecting the x- and the y-axes, but with an amplitude varying from point to point. Find the planes of constant amplitude.

Answer to Problem 9-2

HINT: The excess pressure for the superposed wave is given by $p(x, y, z, t) = p_1 + p_2 = 2A \cos(k \frac{x-y}{2}) \cos(k \frac{x+y}{2} - \omega t)$. This is a plane wave with wave vector $\frac{k}{2}(\hat{i} + \hat{j})$, i.e., the direction of propagation is parallel to $(\hat{i} + \hat{j})$, which bisects the x- and y-axes. The surfaces of constant phase are planes perpendicular to this direction. However, the amplitude of the wave, being given by the expression $A' = 2A \cos(k \frac{x-y}{2})$, varies from point to point. The surfaces of constant amplitude are given by $x - y = \text{constant}$, i.e., these are planes perpendicular to the wave fronts. Such waves, for which the surfaces of constant phase (the wave fronts) differ from those of constant amplitude, are referred to as *inhomogeneous* ones.

9.7.3 Sources and wave patterns: summary

In summary, sources of sound may be simple or relatively complex ones, where each source has its own characteristic distribution of excess pressure and intensity. A monopole source generates a spherical wave, with a spherically symmetric distribution of the excess pressure and intensity pattern, which reduces to a plane wave pattern at large distances from the source. Dipole and quadrupole sources, on the other hand, generate intensity patterns with an angular dependence.

A vibrating tuning fork can be described effectively as a linear quadrupole source with a harmonic time variation, for which the near zone pattern differs from the far zone one (there exist other, less commonly observed, modes of vibration of the tuning fork). More generally, the excess pressure distribution for any source can be described as a superposition of the patterns due to sources of different multipolarities.

The term 'spherical wave' needs clarification. Speaking broadly, spherical waves have a hierarchy where they can be classified into monopole, dipole, quadrupole, and of higher multipole types. As I have indicated above, the monopole wave is isotropic in space, i.e., has no angular dependence. Spherical waves of higher multipolarities all have an angular dependence, though the angular dependence in the far zone is simpler than that in the near zone. In this book, I use term spherical wave mostly in the sense of the monopole one. A wave of a general description can be described as a superposition of waves of various multipolarities.

Other than plane waves and spherical waves (of various multipolarities, with characteristic angular distributions), one can have *cylindrical* waves, or waves of even more complex spatial distributions of excess pressure and intensity.

The presence of obstacles or other bodies in the surrounding medium of a source alters the distribution of excess pressure, which continues to be described by the wave equation, but the solution describing the excess pressure distribution or, equivalently, the pattern of sound propagation now depends on a set of boundary conditions at the surfaces of these obstacles.

In this book, I will not go into detailed considerations of how to determine the excess pressure distribution for a given disposition of obstacles around a source, which in itself is a difficult problem to address. Instead, we will have a brief look at a number of phenomena of a general nature that arise when a plane wave encounters such an obstacle. Of special importance in this context are the phenomena of *reflection and refraction* which take place when a wave (which we will commonly assume to be a plane wave) encounters a surface separating two homogeneous media.

The phenomena relating to acoustic waves outlined here are observed, in a general way, for waves of other physical descriptions as well, such as *electromagnetic waves* considered in chapter 14.

9.8 Reflection and refraction of plane waves

Fig. 9-11 shows a plane boundary surface separating two distinct media A and B. Imagine a monochromatic plane progressive wave to be incident on this surface from the medium A. This will then give rise to a *reflected* and a *refracted* wave propagating along two different directions in A and B respectively. In the figure, PO represents the wave normal for the incident wave, while OQ and OR correspond to reflected and refracted wave normals respectively. Wave fronts are represented by short line segments perpendicular to the wave normals.

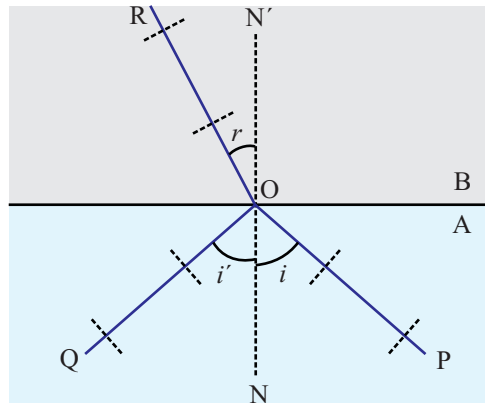


Figure 9-11: Reflection and refraction of a plane wave at the plane interface of two media (A and B); an incident ray (wave normal for a plane progressive wave) PO gives rise to a reflected ray OQ and a refracted ray OR; $N'ON$ is the normal to the interface at the point of incidence; the angles i , i' and r are related to one another by the laws of reflection and refraction; successive wave fronts in the incident, reflected, and refracted waves are shown as sets of parallel line segments, to which the respective wave normals are perpendicular.

Digression: Geometrical wave fronts and rays.

The path PO in fig. 9-11 is often referred to as a *ray* incident at O, while OQ and OR are referred to as the reflected and the refracted rays respectively. The concept of a ray

is often a convenient means to describe the propagation of waves. The rays are straight lines for plane and spherical wave fronts, and are bent or curved lines for waves of other description. However, though a convenient concept in practice, the idea of rays is nevertheless a simplification. This simplified concept is found to be relevant for light waves, which are electromagnetic waves of *small* wavelength (see chapters 14 and 15), in numerous situations of interest, while it is often not relevant for sound waves, the latter being elastic waves of comparatively *larger* wavelength.

The exact description of wave motion in terms of solutions of the wave equation subject to appropriate boundary conditions is, in general, a question of quite considerable complexity. For instance, a description in terms of the motion of wave fronts is not of general validity since it assumes a wave solution that can be expressed in terms of an amplitude and a phase, where the phase varies with time much more rapidly compared to the amplitude.

Such a decomposition into a rapidly varying phase and a slowly varying amplitude is not always possible. Even when one can make such a decomposition, it may not be possible to calculate the phase exactly by solving the wave equation. However, in numerous situations of interest, an approximate calculation of the phase can be carried out and a description of wave motion in terms of approximate wavefront-like surfaces is possible. Such approximate wave fronts are referred to as *geometrical wave fronts*. Corresponding to the geometrical wave fronts one can define a set of paths normal to these wave fronts, along which there occurs the *transport of energy* resulting from the wave motion.

It is such an approximate description of wave motion that is referred to as the *ray* approximation. In the present chapter I will, at times, make use of these concepts relating to rays and geometrical wave fronts without always attempting to analyze how far these are applicable in the strict sense of the term. In the illustrations, for instance, sets of wavefront-like surfaces will be depicted that can be interpreted as geometrical wave fronts, and these will be used only to explain qualitatively a few general features of motion of waves past obstacles and interfaces. Conclusions drawn from such considerations may not be rigorously valid, especially for those regions where the geometrical

wave fronts turn out to be sharply bent or curved.

Laws of reflection and refraction.

The directions of wave normals for the incident, reflected, and refracted waves (as I have mentioned above, I will refer to these as ‘rays’ or ‘ray paths’ in the present context) are found to be related by means of a number of definite rules. These are termed the *laws of reflection and refraction*. In the figure one finds that the incident ray, reflected ray, refracted ray, and the normal (NON') to the boundary surface at the point of incidence O, *all lie in the same plane* (the plane of the paper in this instance). This describes one of the laws of reflection and refraction. Looking at the angles i , i' and r in the figure, we refer to these as the angles of incidence, reflection, and refraction respectively. The second set of laws of reflection and refraction can then be stated as

$$i = i', \quad \frac{\sin i}{\sin r} = \frac{v_1}{v_2}, \quad (9-17)$$

where v_1 and v_2 stand for the phase velocities of the wave in medium A and B respectively. Note that the direction of propagation of the wave undergoes a deviation in refraction due to the difference if the velocities of propagation of the wave in the two media.

1. The phase velocity of a plane progressive monochromatic wave in a medium depends, in general, on the frequency of the wave, as a result of which the bending of the wave in refraction, as described by the second equality in eq. (9-17) differs for waves of different frequencies. This phenomenon is referred to as *dispersion*. For acoustic waves of small amplitude in air or any other fluid, however, the frequency dependence of the phase velocity is negligible.
2. As I have mentioned above, the wave function (i.e., excess pressure in the case of sound waves) has to satisfy a set of *boundary conditions* at the surface separating two media. These boundary conditions imply that the phase has to vary *continuously* across the boundary surface, i.e., in other words, the phase has to approach the *same* value as any given point in the surface is approached from either side of the surface in the two respective media. Denoting the position vector of the point

on the surface by \mathbf{r} , and invoking equation (9-12), one then obtains

$$\mathbf{k}_1 \cdot \mathbf{r} = \mathbf{k}_2 \cdot \mathbf{r} = \mathbf{k}_3 \cdot \mathbf{r}, \quad (9-18)$$

where \mathbf{k}_1 , \mathbf{k}_2 , \mathbf{k}_3 denote the three wave-vectors. One can then arrive at the above laws of reflection and refraction from these boundary conditions. In other words, the laws of reflection and refraction are, in a sense, a consequence of the boundary conditions at the surface of separation.

In the above paragraphs I have assumed the plane interface between the two media to be smooth and infinitely extended. In practice, an interface that a wave encounters in the course of its progress may be curved, and may also be rough, with undulations or protuberances giving it an uneven appearance while, in addition, the interface is necessarily of a finite extent. It is not a simple matter to state what fate awaits a wave as it encounters such a real life interface since the simplified picture based on the laws of reflection and refraction may break down, requiring a more involved description.

The phenomena of reflection and refraction will be considered at greater length in the context of *optics* in chapter 10.

9.9 Diffraction and scattering by obstacles

9.9.1 Finite extent of interface

Let me take up the issue of the finite extent of the interface first. If s denotes the typical or characteristic linear dimension of the interface, then it can be *effectively* taken to be infinite, if it is large compared to the wavelength (λ) of the wave under consideration. On the other hand, if it is not so large compared to λ , then the interface acts effectively like an *obstacle* in the path of the wave, as in fig. 9-12. Assuming the interface to be smooth, its effect on the wave can be described in accordance with the laws of reflection and refraction for regions away from the edges (incidentally, for sound waves, the refracted wave is often not of much consequence since it is absorbed in the second medium).

Close to the edges, however, the wave fronts *get bent* as the wave spreads around the edges and enters into the shadow region, i.e., the region where no wave disturbance can reach according to the ray description. Such bending of a wave around the edges of an obstacle is, under certain circumstances, referred to as *diffraction*.

Figure 9-13 illustrates the bending effect as a wave moves through an aperture in an extended wall or screen, where the edges of the wall act as the obstacles.

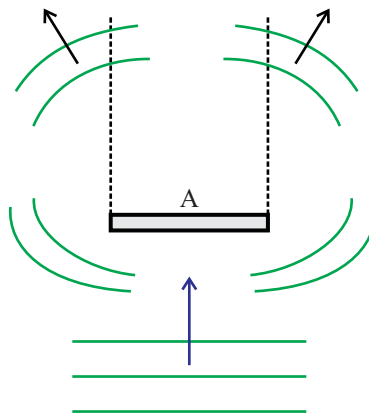


Figure 9-12: Bending of wave around the edges of an obstacle (A); a plane wave incident on the obstacle gets bent around the edges and enters into the shadow region whose boundary is indicated with dotted lines; parts of successive incident plane wave fronts and pieces of the diffracted or bent wave fronts are shown schematically, along with arrows to indicate direction of propagation.

The theory describing such bending is somewhat more general and more involved compared to the description of wave propagation in terms of the ray picture, and works quite well when the extent of the obstacle is several times (roughly, 10 to 100 times) the wavelength. For obstacles of smaller size, the wave fronts break up into fragments that do not join into regular shapes, different parts of the wave being tossed around in different directions by the obstacle, somewhat as in fig. 9-14. Such a process of breaking up of a wave is referred to as *scattering*. However, though the wave assumes a broken appearance in regions close to the obstacle, it reassembles into regular shapes *far away* from it, as seen in the figure. This fact is made use of in making up a theory of the scattering process.

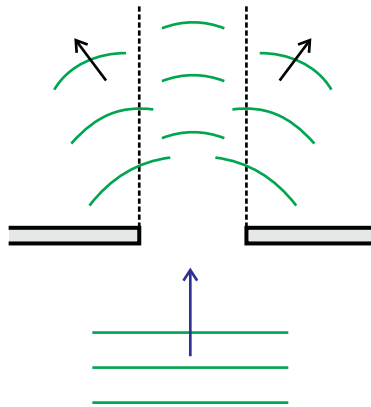


Figure 9-13: Spreading of wave past an aperture in an extended wall; successive wave fronts of the incident plane wave are shown, along with an arrow indicating the direction of wave normal; the wave is diffracted around the edges of the wall and enters into the shadow region, whose boundary is shown with dotted lines; pieces of the diffracted wave fronts are shown schematically, with arrows indicating directions of propagation (the ray paths in an approximate description).

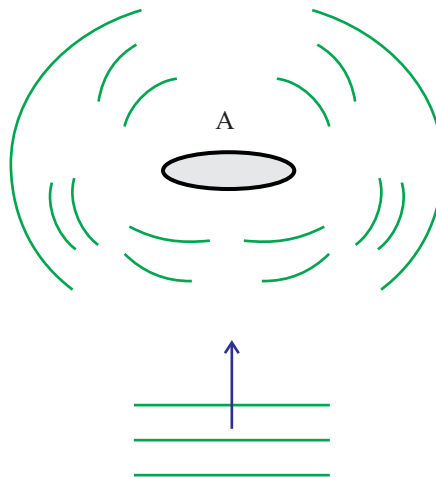


Figure 9-14: Scattering of wave from an obstacle (A); successive wave fronts of an incident plane wave are shown, with the arrow indicating the wave normal; parts of wave fronts scattered in all directions are shown; far from the scatterer, the wave fronts are nearly spherical.

9.9.2 Curvature of the interface

Fig. 9-15 depicts a plane wave incident on a curved but smooth interface of a large extent ($s \gg \lambda$), the latter implying that diffraction effects near the edges of the interface are of no consequence. If the curvature of the surface is small (i.e., the *radius of curvature* is large compared to λ) then the ray picture happens to be applicable once again, and the

laws of reflection and refraction of rays can be invoked to describe the propagation of the wave. What makes this situation different from the one involving a plane interface, however, is that, because of the curvature, the direction of propagation of the wave gets bent in different ways at different points on the interface. As a consequence, a bunch of parallel rays may get converted into a convergent or a divergent bunch by the interface. This fact is made use of in the construction of *lenses* in optics. On the other hand, curvature effects are, commonly, of not much relevance in the case of acoustic waves.

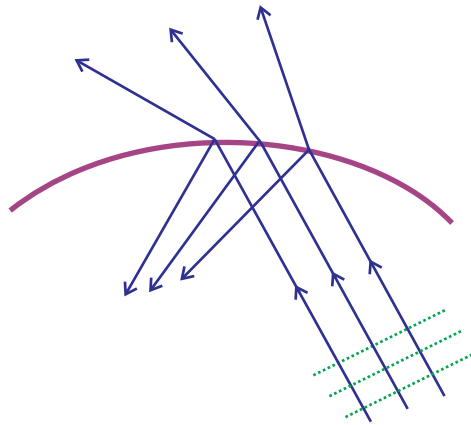


Figure 9-15: Regular reflection and refraction from a curved surface; dotted lines show successive wave fronts of the incident plane wave, while solid lines with arrows depict incident, reflected, and refracted rays; convergent or divergent bunches are produced from a parallel bunch of incident rays.

9.9.3 Wave incident on an uneven surface

An interface or obstacle in the path of a wave can, from a general point of view, be described as an *inhomogeneity*: the wave encounters a break in uniformity of the medium through which it propagates. What is important here is the *length scale* of the inhomogeneity, i.e., the linear extent of an inhomogeneity and, if several inhomogeneities are involved, the typical distance between successive ones. Depending on the length scale as compared to the wavelength, one or more of the basic processes introduced above, i.e., regular reflection (and refraction), diffraction, and scattering may be involved in the way the wave negotiates the obstacle.

Fig. 9-16 shows a plane wave incident on an uneven surface with protrusions giving it a wavy or corrugated appearance. In 9-16(A) the radius of curvature of a corrugation is large compared to the wavelength of the wave incident on it, and regular reflection (and refraction) takes place, that can be described in terms of rays. The rays reflected from one corrugation are oriented differently compared to those reflected from a different one, and all these rays then make up a complex pattern. This is referred to as *diffuse* reflection. In 9-16(B), on the other hand, the protrusions are of a smaller extent, and the wave is diffracted and scattered from these, making up once again a complex pattern.

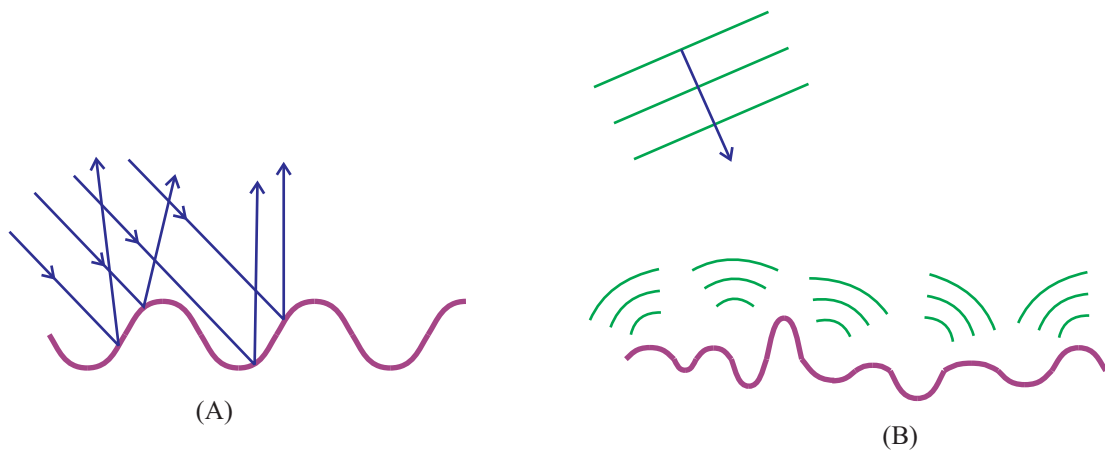


Figure 9-16: Wave incident on an uneven surface: (A) diffuse reflection from corrugations of dimension large compared to wavelength; the reflected rays make up a complex pattern (B) diffraction and scattering from corrugations of dimension comparable to or smaller than the wavelength.

At a basic level, an obstacle or an inhomogeneity may be looked upon as being made up of small parts or *elements* where each element may comprise of an atom or a group of atoms, or even may be of macroscopic, though small, dimensions. Each of these elements scatters the wave independently of the others, and the resultant effect of the inhomogeneity is obtained by *superposition* of all these scattered waves. Looked at from this point of view, even regular reflection and refraction may be interpreted as the result of superposition of a large number of scattered waves.

Since the effect of an inhomogeneity depends on the length scale of the inhomogeneity relative to the wavelength, waves of different kinds are affected in widely different ways

by an inhomogeneity. A beam of light made to be incident on a tree trunk gets diffusely scattered from the rough surface, while a substantial portion is absorbed. On the other hand, a sound wave, being of much larger wavelength, bends around the trunk and proceeds to the other side of the tree with relatively small loss by scattering and absorption.

9.10 Echo and reverberation of sound

9.10.1 Echo

The term ‘echo’ stands for reflected sound that is heard by a listener at a certain time interval after the direct sound is heard by her, the time interval being due to the extra distance travelled by the reflected wave as compared to the direct wave.

Fig. 9-17 illustrates the phenomenon of echo where the direct sound and the reflected sound are depicted with the help of rays since a description in terms of rays is to a certain extent a valid one in respect of the formation of echos. Let the path length for the direct ray be l , and the total path length for the reflected ray be l' . The time interval between the direct sound and the echo is then

$$\tau = \frac{l' - l}{v}, \quad (9-19)$$

where v stands for the velocity of sound. This time interval is of considerable relevance since if it is too small then it cannot be heard as a distinct sound by the listener. After the direct sound reaches the listener, a certain minimum time interval is to elapse if the echo is to be registered distinctly by her since there is a *persistence time* (~ 0.1 s) during which the sensation of hearing a sound continues in the perception of the listener. Denoting this persistence time as τ_p , the condition for hearing an echo distinctly is given by

$$\tau > \tau_p, \quad (9-20)$$

where τ is the time interval given by eq. (9-19).

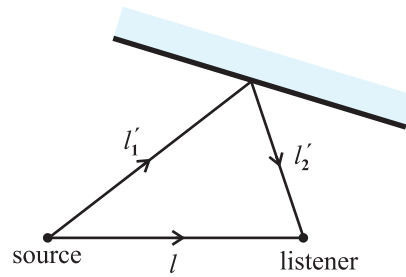


Figure 9-17: Ray diagram for the formation of echo; the direct path length from the source (S) to observer (O) is l , while the path length for the reflected sound is $l' = l'_1 + l'_2$; the difference $l' - l$ is to be sufficiently large for the echo to be heard distinctly.

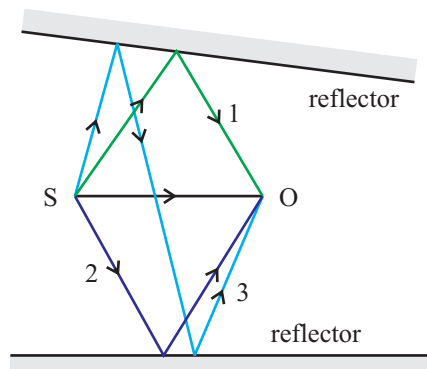


Figure 9-18: Ray diagram illustrating multiple echo in the presence of two reflectors; in addition to the direct ray, three other ray paths are shown (schematic).

Fig. 9-18 illustrates the phenomenon of *multiple echo* where, after hearing the direct sound, the listener hears more than one echos, due to reflected sound reaching her through more than one routes by reflection. Once again, the time interval between two successive echos has to be greater than τ_p , as the interval between the direct sound and the first echo has to be. In the figure, the first and the second echos, corresponding to ray paths marked '1' and '2', involve one reflection each while the third echo (path '3') involves two successive reflections.

Echos are commonly heard at places like those near hills, large buildings and walls. If the extent of a wall, for instance, is large compared to the wavelength of sound, acoustic waves undergo regular reflections from it even though the surface of the wall may be

a rough one, since the length scale of the roughness is usually small compared to the wavelength.

9.10.2 Reverberation.

Reverberation is a phenomenon involving multiple reflection of sound in an enclosed space, where the time intervals between hearing the sound from successive reflections is *less than* the persistence time τ_p so that, instead of hearing the sound from these reflections distinctly, the listener hears an indistinct and persistent rolling sound. The best way to understand what happens in reverberation is to create a single sharp sound in an enclosed space which first creates the impression of a similar sharp sound for the listener as the direct wave reaches her. After that, another short interval may ensue before the reflected sound reaches her for the first time. Thereafter, the reverberation sets in, when the listener has the impression of hearing an indistinct, fluctuating, and persistent sound for some time. Eventually, the reverberation dies down as the sound is absorbed by the walls of the enclosure and by other objects in it.

Reverberation of sound is a complex phenomenon that may be analyzed from the ray point of view as also from the wave point of view, the two being, of course, complementary to each other. A ray proceeding from a point in an enclosed region of space may go on being reflected from the boundary walls of that region without ever following a repetitive course, and all the rays taken together (see fig. 9-19) may form a complex pattern of energy flow in the enclosure, forming the sensation of an irregularly fluctuating persistent sound for the listener located anywhere in the enclosure.

Expressed differently, an initial wave disturbance created in the enclosure, say the one generated by a single sharp sound at any given point, leads to a distribution of the excess pressure within it that goes on evolving in time whereby eventually a fluctuating pressure distribution is set up throughout the enclosed region under consideration. Such a complex pattern of excess pressure distribution differs from a regular standing wave (see sec. 9.15.2) in its complexity and in the fluctuations associated with it.

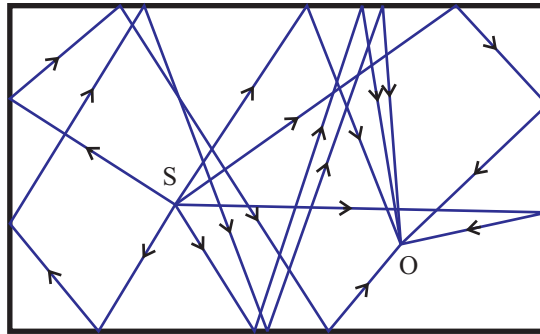


Figure 9-19: Ray diagram (schematic) illustrating the occurrence of reverberation; multiple ray paths are shown from a source point S to a point O where a listener may be located.

Fluctuating wave patterns of similar complexity are formed with electromagnetic waves in an enclosure as well. It appears that the formation of such complex patterns depends on the *shape* of the boundary surface of the enclosed region under consideration. For instance, while regular standing waves may be formed in a rectangular enclosure, complex wave patterns emerge for an oblong *stadium shaped* enclosure.

Reverberation is of crucial importance, for instance, in *auditorium acoustics*, which deals with ways for improving the acoustical qualities of auditoriums where a large number of listeners are distributed throughout an enclosed space in order to listen to the sound created at a particular spot, such as a dais where a performance is being staged.

Too much of reverberation in such an auditorium is detrimental for the success of the performance since it stands in the way of the listeners following the proceedings clearly and distinctly. On the other hand, too little of reverberation is also not desirable since it creates the impression of flat or 'dead' sound. Moreover, reverberation has the desirable effect of evening out the sound throughout the enclosed space under consideration where, in the absence of reverberation, different persons would hear sounds of different intensities since, for a point source, the intensity falls off as $\frac{1}{r^2}$ with the distance from the source (see sec. 9.11.4).

An important quantitative measure of the extent to which reverberation takes place in an enclosure is the *reverberation time*, i.e., the time for which the reverberation persists

when a single sharp sound is created in an enclosed space. It is found to vary in direct proportion to the volume of the space and inversely as the surface area of the boundary walls, being inversely proportional to the mean *absorption coefficient* of the wall surface as well.

small

1. The absorption coefficient measures the fraction of acoustic energy incident on a surface that is absorbed by it.
2. The acoustics of auditoriums and large buildings can have subtle aspects to it. For instance, multiple reflections at the walls of a large room may lead to the production of *standing waves* in the room (see sec 9.15.2). When a sound is created in the room, there may be more than one possible *modes* of distribution of the excess pressure field in it where effects such as reverberation, production of interference patterns (see sec. 9.15.1), and standing waves may come into play.

For instance, in certain halls with vault-shaped domes or specially shaped walls, the curious phenomenon involving the so-called *whispering gallery* mode is observed. When one whispers faintly, close to the wall of the room, it can be heard at a great distance by a listener located near the wall at the diametrically opposite point in the room. This strange phenomenon is the result of a pattern being generated by multiple reflections as the acoustic wave grazes along the walls of the room. However, a complete explanation of this and other subtle acoustic phenomena involves more complex considerations relating to wave motions in enclosures.

9.11 Velocity, energy density, and intensity

9.11.1 Formulation of the problem

As I have already indicated, a wave is generated when the oscillations of some physical quantity originating in a limited region in space from some source gets transmitted to

adjacent regions and then to farther regions away from the source. The nature of the wave depends crucially on two sets of conditions, namely, a set of *initial conditions* relating to how the oscillations originated in time and a set of *boundary conditions* relating to how the physical quantity under consideration is to behave near the surfaces of various objects like reflectors and obstacles present in the path of the wave.

We have found examples of waves of simple as well as of more complex descriptions. Among these, the plane progressive monochromatic wave happens to be of an especially simple nature. Supposing that a wave of this type progresses along the x -axis of a Cartesian co-ordinate system, equation (9-6) is an expression representing the variation of the wave function with respect to x and t .

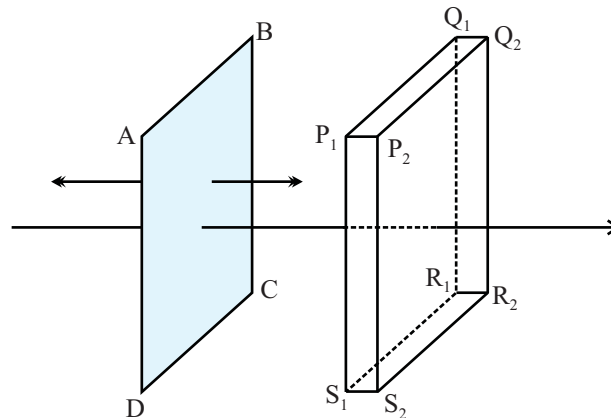


Figure 9-20: Oscillations of the rigid diaphragm ABCD generate a plane monochromatic acoustic wave that sets up stress and strain in every volume element in the surrounding medium; a thin slice of the medium is shown in which pressure and volume alterations take place; the long arrow denotes the positive direction of the x -axis; the instantaneous position of the diaphragm is parallel to the y - z plane.

For the present, we are concerned with acoustic waves which, as we recall, are pressure waves of small amplitude set up in a fluid, of frequencies lying in some definite range, though in the discussion to follow, the limits of the frequency range will be found not to be of much relevance. For the sake of concreteness, you may think of sound waves in air since the waves do usually travel in air before generating the sensation of sound in our ears.

Imagine that a rigid diaphragm, infinitely extended in all directions in the y - z plane, is being made to oscillate along the x -axis with some definite frequency wherein the diaphragm performs a simple harmonic motion in the x -direction. This will cause a monochromatic progressive wave to be generated in the fluid (which we assume to be isotropic, homogeneous, and infinitely extended) on either side of the diaphragm, say, on the side extending up to $x \rightarrow \infty$.

In figure 9-20, ABCD represents the mean position of one portion of the diaphragm. As a result of the oscillations of the diaphragm, the layers of air adjacent to it will alternately be compressed and rarefied. In the figure, $P_1Q_1R_1S_1$ and $P_2Q_2R_2S_2$ indicate the positions at any given instant of time of the boundary surfaces of one such layer, only a portion of the layer in the shape of a rectangular parallelepiped being shown. In the following, we will not concern ourselves with the random thermal motions of the air molecules (see chapter 8) making up small volume elements in the layers, but will only look at the *average* motions of these volume elements. When we speak of motions of air 'particles', we will actually mean such average motions, including deformations, of small volume elements.

The condition under which the random thermal motion of the air molecules can be averaged away is that the mean interval between successive collisions suffered by a molecule in its thermal motion (see sec 8.10.7) be small compared to the typical time period associated with the acoustic wave set up in the medium, which one may take as the lowest time period associated with the wave function at any given point. This condition is indeed satisfied under commonly occurring situations relating to sound waves.

What can one say about the motion of the air particles in the above layer due to the oscillations of the diaphragm (depicted with arrows heading both ways in fig. 9-20)? Evidently, such motion can only be possible along the x -axis, motion in the y - z plane being ruled out by the symmetry of the problem. The oscillations of the diaphragm being simple harmonic in nature, small volume elements in a typical layer like the one shown

in the figure will perform simple harmonic oscillations of two types - an alteration in elastic strain and stress, and an alteration in the position along the x -axis that takes place about the mean position of the element, the two being, moreover, related to each other (see formulae (9-23) below). Of these, we consider first the alteration in the elastic strain, which appears as a simple harmonic variation of the thickness of the layer. As the thickness of the layer varies, the area of the end faces ($P_1Q_1R_1S_1$ and $P_2Q_2R_2S_2$) will remain unaltered (reason out why) and hence, the proportional change of the volume of the layer will be the *same* as that of its thickness (check this out).

As mentioned above, the proportional change in the thickness or volume of the layer under consideration corresponds to an elastic *strain* produced in it. While the proportional change in thickness represents the longitudinal strain along the x -axis, the proportional change in volume corresponds to *bulk* or volume strain. In the situation under consideration the two are equal, say, e .

A longitudinal strain along the x -axis is equivalent to a pure volume strain, together with longitudinal strains along the x -, y -, and z -axes.

9.11.2 Displacement and strain

In chapters 6 and 7, we have seen that the only type of stress that can be generated in a fluid is bulk stress (we ignore the viscosity of the fluid for our present purpose; incidentally, the *strain* in the fluid need not be a pure bulk strain), and that this bulk stress at any given point is nothing but the pressure of the fluid at that point, taken with a negative sign. Thus, if p represents the excess pressure in air due to the setting up of the sound wave, then the corresponding additional volume stress generated will be $-p$. Moreover, the ratio of the volume stress and volume strain being the bulk modulus (K) of the fluid, we have, in the situation under consideration,

$$e = -\frac{p}{K}. \quad (9-21)$$

The term 'bulk modulus' is not a precise one since one has to specify the condition

under which the strain and stress occur. For instance, deformation occurring under *adiabatic* and *isothermal* conditions correspond to distinct values of the bulk modulus.

In the case of acoustic waves, it is the *adiabatic bulk modulus* that is commonly of relevance; see sec. 9.11.3.1.

Formula (9-21) relates the excess pressure due to the sound wave with the excess bulk strain in the fluid (which equals the excess longitudinal strain in the present case; incidentally, in the context of acoustic waves, the bulk strain with a negative sign is referred to as the *condensation*). For instance, if the excess pressure varies with x and t in accordance with, equation (9-4), then the longitudinal strain, or equivalently, the bulk strain, will vary as

$$e = -\frac{p_0}{K} \cos(kx - \omega t). \quad (9-22)$$

Since a fluid, under commonly observed situations, is always characterized by a stress and strain, we will consider here the excess stress and excess strain with reference to the situation where there is no acoustic wave passing through it. With this understanding, I will often do away with the qualifying term ‘excess’ for the sake of brevity.

Thus, instead of describing a sound wave as a pressure wave in a fluid, one can alternatively describe it as a *strain* wave as well. What is more, a number of other physical quantities may also be looked upon as wave functions in the description of a sound wave since these are all related to excess pressure, stress, or strain. For instance, along with the change in thickness of a layer like the one considered in fig. 9-20, there also occurs a change in the mean position of any volume element of the fluid within the layer. Denoting the displacement along the x -axis as s , the instantaneous velocity of the boundary surface will be $u = \frac{\partial s}{\partial t}$, and the instantaneous acceleration will be $a = \frac{\partial^2 s}{\partial t^2}$, which are nothing but derivatives of displacement with time, of respectively the first and second orders, with the co-ordinate x held fixed.

The fact that plane acoustic waves in a fluid can be described in terms of the longitudi-

nal strain and the displacement (s) or velocity (u) along the direction of propagation of the wave (this applies to spherical waves as well) has led to the convention of referring to these waves as *longitudinal* ones. By contrast, electromagnetic waves (see chapter 14) are *transverse* in nature. A more logical way of classifying waves of various different descriptions, however, is to base the classification on the nature of the wave function involved, i.e., taking note of whether the wave function is a scalar, a vector, or a tensor. Looked at from this point of view, sound waves are more logically described as *scalar* ones since the excess pressure at any point of the fluid is a scalar, starting from which the other wave functions like the longitudinal strain and the displacement can be derived.

The description of sound waves as longitudinal ones is not limited to plane and spherical waves alone since, for a wave of a more general type, formed by the superposition of a number of plane or spherical waves, a similar description remains valid. For an acoustic wave of a specified frequency (i.e., a monochromatic wave), one can define a wave front as a surface on which the instantaneous value of the wave function (i.e., the physical quantity whose space-time variation defines the wave) remains constant. Given a point on the wave front, the instantaneous direction of propagation of the wave front in the neighborhood of that point then coincides with the direction of the instantaneous velocity of a small element of the fluid medium centered around it.

9.11.3 Velocity of sound in a fluid

While describing a sound wave one can use any one of the quantities s , u , or a as the wave function just like p or e because all these physical quantities are mutually related. For instance, the longitudinal strain e and the excess pressure are related to the displacement s as

$$e = \frac{\partial s}{\partial x}, \quad p = -K \frac{\partial s}{\partial x}. \quad (9-23)$$

Considering now a layer of the fluid as in fig. 9-20, one can set up its equation of motion by equating its acceleration with the net force per unit mass acting on it in the x-

direction where this net force can be worked out from the difference of excess pressures between the two surfaces bounding it. Making use of the second relation in (9-23) in the resulting formula one ends up with the relation

$$\frac{\partial^2 s}{\partial t^2} = \frac{K}{\rho} \frac{\partial^2 s}{\partial x^2}. \quad (9-24)$$

Comparing with eq. (9-14), one recognizes this as precisely the wave equation in one dimension, for which the parameter v is given by

$$v = \sqrt{\frac{K}{\rho}}. \quad (9-25)$$

This is a good place to pause and take stock of how we arrived at these results. We started by considering a situation, described by fig. 9-20, in which a one dimensional wave is set up in a fluid and we saw that the resulting wave involves variations of the excess pressure and the longitudinal strain (and equivalently, of the bulk strain) with the position co-ordinate x and time t , as also an oscillatory motion of its mean position. By looking at a thin layer of the fluid and setting up its equation of motion, one finds that the latter is equivalent to eq. (9-24), which is of the same form as the one dimensional wave equation (9-14), with v given by eq. (9-25).

As I have mentioned above, the one dimensional wave equation possesses solutions of various different descriptions depending on initial and boundary conditions, a class of solutions of the simplest nature being the plane monochromatic ones. It is for such an assumed solution, given by eq. (9-4), with an angular frequency ω and a wavelength $\lambda = \frac{2\pi}{k}$ that the bulk strain is given by eq. (9-22). The ratio $\frac{\omega}{k}$ for such a wave gives its phase velocity which, by substitution, is seen to be the same as the parameter v in the wave equation. In other words, the phase velocity of a monochromatic plane progressive wave in a fluid is given by the expression (9-25).

The derivation of eq. (9-24) involves a number of assumptions that I have not stated explicitly. One of these is the assumption that the displacement s or the excess pressure p for any value of x and t are to be sufficiently small so that their squares and

higher powers can be ignored without bringing in appreciable errors. A different way of expressing the same thing is to say that the waves described by the equation are *small amplitude* ones. Equations describing waves of larger amplitudes in a fluid are relatively more complex in that these are *nonlinear* differential equations, and the principle of superposition does not apply to such waves.

Another implied assumption is the one that the effects of viscosity are negligible in the propagation of the waves. In reality, the viscosity of the fluid results in a damping of the waves set up in it.

Finally, the waves being assumed to be related to pressure variations in the *bulk* of the fluid, surface tension effects have also been assumed to be of no consequence. Waves set up near the surface of a liquid may depend strongly on its surface tension.

One can also set up the equation describing small amplitude waves propagating along all three directions in a fluid. This is seen to be precisely the three dimensional wave equation (eq. (9-15)) we have already had a look at, with the parameter v once again given by eq. (9-25). As I have mentioned, examples of simple wave solutions to the three dimensional wave equation are the plane monochromatic waves of the general form (9-11) and the spherical wave of the form (9-16), for either of which the velocity is given by $v = \sqrt{\frac{K}{\rho}}$. In addition, innumerable other solutions of a more complex nature are possible, among which are the sound emitted by a dipole or a quadrupole source, or the wave resulting from the presence of one or more obstacles around a source of sound. Additionally, a wave may be characterized by a range of frequencies rather than a sharply defined one. For waves of such general description, the concept of phase velocity may not always be a well defined one. In the case of acoustic waves, however, the phase velocity is by and large seen to be of relevance since these are, to a large extent, *non-dispersive*.

9.11.3.1 Velocity of sound in an ideal gas

If the fluid under consideration be an ideal gas with kg-molecular weight (commonly referred to as the molar mass) M then the relation between the mean density ρ and

mean pressure P is

$$P = \frac{\rho}{M} RT, \quad (9-26)$$

where T stands for the temperature in the absolute scale (check this out, making use of considerations in chapter 8). It may be mentioned here that, similar to the variation of pressure in the fluid around the mean pressure P , there occurs a variation of density as well around the mean density ρ , and one can look upon the proportional variation of density as another wave function describing the sound wave under consideration.

When a monochromatic acoustic wave is set up in the gas, the instantaneous pressure at any given point varies simple harmonically about the mean pressure, and the deviation of the instantaneous pressure from the mean pressure gives the instantaneous value of the excess pressure. Similarly the instantaneous density differs from the mean density in the case of a gas (for a liquid, the proportional variation in density is negligible), where the latter is related to the mean pressure by eq. (9-26).

The bulk modulus K in a gas happens to be related to the pressure P . However, in order to define bulk modulus unambiguously, one has to specify the condition under which stress and strain are generated in the fluid (refer back to sec. 9.11.2). For instance, if the stress is developed under an *isothermal* condition, i.e., if the temperature remains unaltered due to exchange of heat with the surroundings, the bulk modulus is given by $K = P$. On the other hand, if the stress is generated *adiabatically*, i.e., with no heat being exchanged between the gas in any given volume element and its surroundings, the bulk modulus works out to

$$K = \gamma P, \quad (9-27)$$

where γ stands for the ratio of the specific heats of the gas at constant pressure and constant volume (c_p and c_v ; see sections 8.21.2 and 8.21.3).

As sound propagates through air, the compression and rarefaction of the successive layers of air happen to occur so rapidly that little heat can be exchanged between any given volume element and the fluid in the surrounding regions, and one can thus assume with

sufficient accuracy that the process of generation of elastic stress in any such volume element occurs adiabatically. This is related to the frequency range corresponding to sound waves I mentioned earlier. In other words, assuming air to be an ideal gas, the velocity of sound in air works out to

$$v = \sqrt{\frac{\gamma P}{\rho}}. \quad (9-28)$$

This is referred to as *Laplace's formula* for the velocity of sound. In writing this equation, I have assumed air to be a pure gas. One can, in this case, make use of (9-26) to obtain an alternative expression for the velocity:

$$v = \sqrt{\frac{\gamma RT}{M}}. \quad (9-29)$$

The generation and propagation of an acoustic wave through a medium is a *response* of the latter to a disturbance produced in it results in a *no-equilibrium* situation. What is important to note is that, for a *small* deviation from the equilibrium configuration, the velocity of propagation is determined by *equilibrium* values of thermodynamic functions characterizing the medium, as in formulae (9-28) and (9-29).

For a mixture of two ideal gases, both (9-26) and (9-29) require appropriate modifications. For instance, instead of (9-29) the expression for the velocity of sound is given by

$$v = \sqrt{\frac{xc_{p1} + (1-x)c_{p2}}{xc_{v1} + (1-x)c_{v2}} \cdot \frac{RT}{xM_1 + (1-x)M_2}}. \quad (9-30)$$

Here M_1 and M_2 are the molar masses of the two components in the gas mixture, c_{p1} , c_{p2} are their respective molar specific heats at constant pressure, and c_{v1} , c_{v2} the respective molar specific heats at constant volume, while x , $(1-x)$ denote their respective mole fractions in the mixture.

It may be mentioned that if the processes of rarefaction and compression of the air layers were slow ones then these would have occurred as isothermal processes because of the possibility of heat exchange between the layers and their surroundings which

could then take place more rapidly compared to these slow processes of rarefaction and compression, and the velocity of sound would then be given by the formula

$$v = \sqrt{\frac{P}{\rho}}, \quad (9-31)$$

instead of (9-28). In reality, however, the processes of rarefaction and compression in air during the propagation of sound waves are sufficiently rapid (as compared with the rate of heat conduction), and (9-28) happens to be the correct formula for the velocity of sound in preference to (9-31), the latter being referred to as Newton's formula. The validity of Newton's formula requires a high value of the thermal conductivity, which is contrary to the idea of an ideal gas. Such a high value of the thermal conductivity would lead to a correspondingly high rate of energy dissipation in the gas, causing elastic waves generated in it to be damped instantaneously.

It may be appropriate at this place to state that the above considerations, including the statement of the formula (9-28), refers to propagation of small amplitude elastic waves in an *ideal fluid*, i.e., one in which *dissipative* processes such as viscous damping and heat conduction are of negligible relevance.

The equations of motion of a fluid under quite general conditions are based on the following three fundamental considerations, namely, ones involving mass balance, momentum balance, and energy balance. The condition of energy balance, for instance, requires that the total flow of energy of a fluid through the surface of any imagined volume in it (from the interior of the volume to its exterior), has to be equal to the energy flowing into the volume in the form of heat as also in the form of work performed on the fluid in that volume. The resulting equations are, generally, speaking, of a phenomenological nature and involve transport coefficients such as the coefficients of viscosity and thermal conductivity of the fluid, these coefficients themselves being of a phenomenological nature.

Assuming that thermal transport in the fluid is of negligible relevance, one can set up its equation of motion involving its coefficient of viscosity, commonly referred to as the Navier-Stokes equation (refer back to section 7.5.3), which is a non-linear equation,

characterized by the feature that the principle of superposition (refer back to sec. 9.5.3) does not apply to it. If, in addition to heat transfer, viscosity effects are also assumed to be of negligible effect, one is led to the *Euler equation* for the fluid which is still a nonlinear equation of motion (conversely, the assumption of a non-zero viscosity leads one from the Euler equation to the Navier-Stokes equation). If now one makes the assumption that the variations in pressure and density are to remain small during the motion of the fluid, one ends up with a *linearized* equation of motion that leads to the wave equation we encountered in sec. 9.6, with the speed given by (9-25), where K is to be interpreted as the adiabatic bulk modulus. A particularly useful wave function in this context is the *velocity potential* which is a scalar function $\phi(\mathbf{r}, t)$ of position \mathbf{r} and time t , related to the velocity $\mathbf{v}(\mathbf{r}, t)$ as $\mathbf{v} = \text{grad}\phi$. The wave equation satisfied by ϕ describes the propagation of sound waves, which can be described as small amplitude elastic waves in an ideal fluid.

Even though heat transfer is assumed to be of negligible relevance, the propagation of a sound wave involves temperature changes along with changes in pressure and density since the three are related by the *equation of state* of the fluid. Since the fluid is assumed to be ideal, the temperature changes occur adiabatically. In the case of a liquid, however, the temperature changes are relatively small owing to low compressibility compared to a gas, and propagation occurs under an approximately isothermal condition even in the case of a liquid with a non-zero thermal conductivity.

Generally speaking, the propagation of sound in a fluid with non-zero values of the thermal conductivity and viscosity is attended with the process of rapid *attenuation* in which the energy density associated with elastic deformation (refer to sec. 9.11.4, and to formula (9-34) below) decreases by means of thermal and viscous dissipation, and the description of elastic waves can no longer be given in simple terms.

9.11.3.2 Dependence on pressure, temperature, and humidity

One can make use of (9-29) and (9-30) to work out the dependence of the velocity of sound in air on the pressure, temperature, and relative humidity of the atmosphere.

Let us first consider the velocity of sound in *dry* air. This is principally a mixture of oxygen and nitrogen. However, the values of c_p and c_v for oxygen are not much different than those for nitrogen, and the ratio $\gamma \equiv \frac{c_p}{c_v}$ has the value 1.4 for both. The molar masses for these two gases are respectively 32 and 28 (the unit 'kg' is commonly suppressed in expressing molar masses). On making use of the known values of the average mole fractions of these two in air one obtains the value 28.8 (approx) for the quantity $M' \equiv xM_1 + (1-x)M_2$ occurring in (9-30) which one can term the mean molar mass of air. Inserting these values in (9-30) one can work out the velocity of sound in dry air at any given temperature. Notice that this velocity *depends on the temperature alone, and not on the atmospheric pressure*. If T_1 and T_2 be two temperatures on the absolute scale then the ratio of velocities of sound in dry air at these two temperatures is given by the formula

$$\frac{v_1}{v_2} = \sqrt{\frac{T_1}{T_2}}. \quad (9-32)$$

We now consider the velocity of sound in *moist* air. We can think of moist air as a mixture of dry air and water vapour. If x be the molar fraction of water vapour in the moist air, then one can again make use of equation (9-30) to work out the velocity of sound in the moist air. We use, in this equation, appropriate values of M_1 ($= 18$), the molar mass of water vapor, and M_2 ($= 28.8$), the mean molar mass of dry air, as also measured values of c_{p1} , c_{v1} , the two molar specific heats of water vapour and c_{p2} , c_{v2} , the corresponding molar specific heats of dry air. Denoting the partial pressure of water vapor in air as f and the total pressure as P , an alternative expression for x is seen to be

$$x = \frac{f}{P}. \quad (9-33)$$

Notice from formula (9-30) that the humidity (expressed as x here) affects the speed of sound on two counts - first, due to the difference in the specific heats of water vapour (c_{p1} , c_{v1}) compared to those (c_{p2} , c_{v2}) of dry air, and second, due to the difference in their molar masses or vapor densities. The first of these two factors causes a *decrease* in the speed of sound with increasing partial pressure of water vapor at constant atmospheric pressure and temperature, while the second one causes the speed of sound to *increase*.

On balance, the speed of sound in moist air increases with increasing humidity.

9.11.4 Energy density and intensity

We will now briefly look at the energy changes associated with the compression and rarefaction of air layers caused by the propagation of a sound wave. Think of the layer shown in fig. 9-20. We have seen that the pressure of the layer oscillates about the mean pressure (P) and, at the same time, there is set up a harmonically varying longitudinal strain as also a bulk strain in the layer. The excess pressure (p) and the strain (e) are both consequences of the elastic deformation of the layer (the bulk stress resulting from the deformation being $-p$). This involves a change in the elastic deformation energy of the layer under consideration. At the same time, the kinetic energy of the layer also undergoes a change as its mean position executes simple harmonic oscillations, along with the oscillations in its thickness.

These energy changes in the layer occur along with *exchange of energy with the surrounding layers* of air by way of the layers performing *work* on one another (however, one can assume, to a good degree of approximation, that no *heat exchange* takes place between the adjacent layers). In the *steady state* when a plane progressive wave of constant amplitude propagates through the air, as much energy enters into the layer under consideration in a given time interval as the energy flowing out of it, and the sum of the energy due to elastic deformation and the kinetic energy of the layer remains constant.

One can thus assume that there is no energy dissipation in the layers of the fluid medium due to the propagation of the acoustic wave. Indeed, the dissipation of energy due to viscous friction and thermal conduction can be ignored for all practical purposes.

This total energy of the layer per unit volume, or its *energy density*, works out to

$$E = \frac{p_0^2}{2K} = \frac{p_0^2}{2\rho v^2}, \quad (9-34)$$

where ρ is the mean density of the layer. It may be mentioned here that both the

strain energy density and the kinetic energy density for a monochromatic wave vary sinusoidally with time and strictly speaking, only the energy densities averaged over sufficiently large time intervals are meaningful. However, for a sinusoidal variation of the two energy densities, the *sum* of the two remains constant and the time averaging is not necessary. Generally speaking, though, a time averaging is implied while referring to values of physical quantities that vary rapidly with time. For any time dependent quantity, say, $f(t)$ (which may additionally depend on spatial co-ordinates; the latter however, are to be held fixed when working out the time average;) the time average over an interval τ is defined as

$$\langle f(t) \rangle = \frac{1}{\tau} \int f(t) dt, \quad (9-35)$$

where the integration is to be performed over an interval of duration τ . In the case of a simple harmonic variation in time, τ can be taken to be the time period of the variation while, more generally, τ is to be taken as an infinitely large interval.

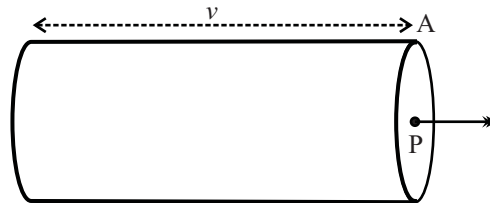


Figure 9-21: Cylindrical volume element of length v with an end face A around the point P; illustrating the flow of energy in an acoustic wave.

As mentioned above, the kinetic energy and the elastic strain energy of any layer of air go on increasing and decreasing due to exchange of energy with surrounding layers. In other words, there occurs a *flow* of energy through the successive layers. The rate of flow of energy can be worked out by referring to fig. 9-21 where P is any point in the medium (which can be assumed to be air for the sake of concreteness) in which a plane monochromatic wave propagating along the x -axis is set up and A is a small area imagined in the y - z plane around P .

Consider a cylindrical volume imagined to be drawn with its axis parallel to the x -axis and with A as one of its end faces (refer to fig. 9-20 for comparison where a volume element of a different shape was considered), the length of the cylinder being v , the phase velocity of the propagating wave. If a denotes the area of the end face then the volume of the cylinder will be av , and multiplying this with the energy density (equation (9-34)) one arrives at the acoustic energy (i.e., the energy due to the sound wave set up in air) contained in the cylindrical volume. Since the length of the cylinder is v , this must represent the amount of energy flowing per unit time through A in the direction of the double headed arrow in the figure, which is thus given by the expression Eva where, as mentioned above, a time averaging is implied while referring to the energy density. *The rate of energy flow per unit area* is then Ev , which is termed the *intensity* of sound. making use of the formula (9-34), the expression for the intensity works out to

$$I = \frac{p_0^2}{2\rho v}. \quad (9-36)$$

Notice that the intensity is proportional to the square of the amplitude of variation of excess pressure and inversely proportional to the mean density as also to the velocity of sound.

In this context the term *energy flux* is relevant. Imagining a surface of area, say, A in a region through which there occurs a flow of acoustic energy, the rate of flow of energy through the area is termed the *flux* through that area.

More generally, considering the flow of any physical quantity of interest, which may be a scalar or a vector (or even a tensor), the rate of flow through any given area is referred to as the *flux* of that quantity through that area.

Though the expressions (9-34) and (9-35) for the energy density and the intensity have been written down with reference to a plane progressive wave, these can be generalized within limits to the case of waves for which the curvature of the wave fronts is small, which turns out to describe a wave at a sufficiently large distance from the source. For instance, these can be employed in the case of a spherical wave produced by a point

source, provided that the spatial variation of p_0 , the amplitude of the excess pressure, is taken into account. For such a wave, the intensity varies inversely as the square of the distance from the source (see sec. 9.11.4.1 below), which means that p_0 varies inversely as the distance.

9.11.4.1 Spherical waves: the inverse square law of intensity.

Fig. 9-22 depicts a source of sound S (imagined to be a point) and two points of observation, P_1 , P_2 , located on a line SX in any given direction, at distances r_1 , r_2 respectively from S . Imagining a narrow cone (dotted lines in the figure) with axis SX , let δW be the energy emitted per unit time by the source within this cone, i.e., in directions close to SX within the limits of the cone. The figure shows two sections of the cone by surfaces passing through P_1 and P_2 , where these surfaces are perpendicular to the axis SX . If the areas intercepted by the cone on these surfaces be δS_1 and δS_2 respectively, then one has

$$\frac{\delta S_1}{r_1^2} = \frac{\delta S_2}{r_2^2}. \quad (9-37)$$

This relation can be established by considering the geometry of the cone (check this out). If the *solid angle* of the cone be $\delta\omega$, then both sides of the above formula are equal to $\delta\omega$ (see sec. 11.8.2.2 for a brief introduction to the concept of solid angle).

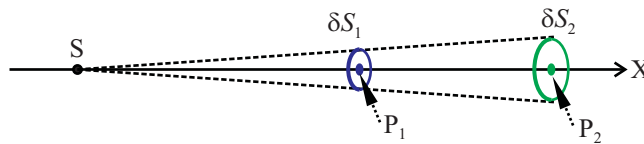


Figure 9-22: Explaining the inverse square law of intensity, with a point source S and two observation points P_1 , P_2 on a line SX in any given direction; the dotted lines represent a narrow cone with axis SX ; δS_1 , δS_2 are areas intercepted by the cone on surfaces through P_1 , P_2 perpendicular to SX ; in the steady state, the rates of flow of acoustic energy through the two areas are the same.

In a steady state, there does not occur any energy accumulation anywhere in the

medium (i.e., the energy density remains constant in time at all points), and the energy δW carried per unit time by the sound from the source within the cone under consideration crosses both the areas δS_1 and δS_2 in unit time. In other words, the intensities I_1 and I_2 at P_1 and P_2 are given by

$$I_1 = \frac{\delta W}{\delta S_1}, \quad I_2 = \frac{\delta W}{\delta S_2}, \quad (9-38)$$

which immediately gives, in view of (9-37),

$$\frac{I_1}{r_1^2} = \frac{I_2}{r_2^2}. \quad (9-39)$$

This is the *inverse square law* of intensity: the intensity at a distance r from the source in any given direction is inversely proportional to r^2 .

In particular, if the sound energy is emitted *isotropically* (i.e., equally in all directions) by a point source then, in the steady state, the intensity at a point at a distance r from the source will be $I = \frac{W}{4\pi r^2}$ regardless of the direction along which the point is located, where W stands for the total rate of emission of sound energy from the source (reason this out).

The unit of intensity is $\text{J}\cdot\text{s}^{-1}\cdot\text{m}^{-2}$, or equivalently, $\text{W}\cdot\text{m}^{-2}$. A common practice in referring to the intensity is to express its value in comparison with that of a standard source of sound kept at a specified distance, or with some standard intensity specified in some other way. This *relative intensity*, expressed in a *logarithmic scale*, is then used as a measure of the intensity level of sound and is expressed in *bel* or *decibel*. If the standard intensity level be I_0 then the intensity expressed in bel and decibel units will be, respectively,

$$B = \log\left(\frac{I}{I_0}\right), \quad (9-40a)$$

and

$$D = 10 \log\left(\frac{I}{I_0}\right). \quad (9-40b)$$

Problem 9-3

The amplitude of acoustic pressure variation at a point 2m away from a source is 0.02Pa. If the density of air be $1.2\text{kg}\cdot\text{m}^{-3}$, and the velocity of sound in air be $350\text{ m}\cdot\text{s}^{-1}$, what is the sound intensity at that point, and what is its decibel level as compared to an intensity $I_0 = 10^{-12}\text{W}\cdot\text{m}^{-2}$? Assuming that the source emits sound isotropically, what is the power output of the source?

Answer to Problem 9-3

HINT: Using formula (9-36), one obtains $I = \frac{(0.02)^2}{2 \times 1.2 \times 350}$ (in SI units), i.e., $4.76 \times 10^{-7}\text{ W}\cdot\text{m}^{-2}$. The decibel level is then $D = 10 \log\left(\frac{4.76 \times 10^{-7}}{10^{-12}}\right) = 10 \times (5 + \log 4.76) = 56.8$ (approx). Since the energy emitted by the source per unit time gets uniformly spread over an area $4\pi r^2$ at a distance r , the rate of energy emission is $W = 4\pi r^2 I$, where I is the intensity at a distance r . This gives $W = 4\pi \times (2)^2 \times 4.76 \times 10^{-7}\text{ W}$, i.e., $2.4 \times 10^{-5}\text{W}$ (approx).

Problem 9-4

A plane progressive monochromatic acoustic wave of intensity $4.0 \times 10^{-7}\text{W}\cdot\text{m}^{-2}$ is set up in a gaseous medium for which the ratio of the two specific heats is $\gamma = \frac{5}{3}$. If the mean pressure in the medium be $P = 1.01 \times 10^5\text{Pa}$, and the mean density is $\rho = 1.2\text{kg}\cdot\text{m}^{-3}$, what is the amplitude of variation of the excess pressure and of the condensation (refer to sec. 9.11.2), which is defined as the bulk strain with a negative sign?

Answer to Problem 9-4

HINT: making use of formulae (9-28) and (9-36), the expression for the intensity can be written as $I = \frac{p_0^2}{2\sqrt{\gamma P \rho}}$, which gives $p_0 = (2I)^{\frac{1}{2}}(\gamma P \rho)^{\frac{1}{4}}$. Substituting the given values, the amplitude of the excess pressure is obtained as $p_0 = (2 \times 4.0 \times 10^{-7})^{\frac{1}{2}} \times (\frac{5}{3} \times 1.2 \times 1.01 \times 10^5)^{\frac{1}{4}}\text{Pa}$, i.e., $18.96 \times 10^{-3}\text{Pa}$ (approx). Again, the expression for the condensation is obtained from formulae (9-21) and (9-27)

as $\epsilon = -e = \frac{p}{\gamma P}$, from which the amplitude of variation of condensation is obtained as $\epsilon_0 = \frac{p_0}{\gamma P} = \frac{18.96 \times 10^{-3}}{\frac{5}{3} \times 1.01 \times 10^5} = 1.13 \times 10^{-7}$ (approx).

9.12 Ultrasonic waves

The term *ultrasound* or ultrasonic waves refers to pressure waves (or, more generally, elastic waves set up in a medium) of frequency larger than the upper limit of the range corresponding to acoustic waves or audible sound. Though the range differs from person to person, waves of frequency greater than 30 KHz are generally considered to qualify as ultrasound.

The basic nature of ultrasound does not differ fundamentally from that of acoustic waves. What makes the ultrasonic waves special is the relatively high frequency and correspondingly small wavelength characterizing these, as compared to the acoustic waves of relatively larger wavelengths (recall that the product of the frequency and the wavelength equals the phase velocity (eq. (9-9)); the latter can be assumed to be a constant for a given medium for small-amplitude waves). As a result, ultrasonic waves differ from the acoustic waves in the way these waves behave when they encounter obstacles in their path.

For instance, while an acoustic wave may bend around and propagate past an object in its path because of its wavelength being large compared to the linear dimensions of the latter, an ultrasonic wave, because of its smaller wavelength, may get reflected from the surface of the object in a more or less regular manner (see sec. 9.8). This makes the ultrasonic waves useful as *probes* for *sensing* and *imaging* purposes.

Ultrasound imaging is a major diagnostic technique in present day medical technology. Another major area of application of ultrasonic waves is in *non-destructive testing* of materials. For instance, when an ultrasound wave is sent into a metal or any other crystalline material, it gets reflected from the *faults* within the material, these being regions in it where the arrangement of the atoms and molecules deviates from a regular

one. The location and nature of the faults can be determined by recording the reflected or diffracted waves.

One major difference of a fundamental nature distinguishing ultrasonic waves from acoustic ones relates to the fact that, because of their smaller wavelengths, ultrasonic waves are commonly characterized by a relatively large amplitude-to-wavelength ratio. The differential equations describing these waves differ from the wave equation (9-15) in that the equations for ultrasonic waves often include *non-linear* terms. These nonlinear terms introduce a number of special features in the propagation of the ultrasonic waves and in their behavior as they encounter obstacles in their path. For instance, the wave profile of an ultrasonic wave may keep on changing as the wave propagates through the medium so that a monochromatic wave with a sinusoidal profile, for instance, may acquire a saw-tooth shape and become rich in harmonics (i.e., multiples of the basic frequency characterizing the wave).

Nonlinear acoustic and ultrasonic waves have practical applications ranging over a wide area, including those in ultrasound imaging and in *levitation* produced by sound waves, the latter being of major use in the electronic microchip industry.

9.13 Döppler effect

9.13.1 Introduction

Consider a point source of sound at rest, emitting spherical waves of angular frequency ω . The variation of excess pressure for such a wave is given by an equation of the form (9-16), which is dominated by the variation of the phase

$$\Phi = kr - \omega t. \quad (9-41)$$

The variation of the excess pressure with time at any given point fixed in space occurs due to the variation of the phase which changes with time at the constant rate ω . One can visualize the process of wave propagation by imagining a succession of spherical wave fronts, each differing from the next by a constant phase, where all the wave fronts

keep on expanding in radius at the constant rate v . In particular, the wave fronts corresponding to phases $\Phi = 2n\pi$, ($n = 0, \pm 1, \pm 2, \dots$) are termed ‘crests’ and those with phases $\Phi = (2n + 1)\pi$, ($n = 0, \pm 1, \pm 2, \dots$) are referred to as ‘troughs’. Considering any fixed point in space, the excess pressure at that point attains a maximum value when successive crests arrive at it, while the arrival of the troughs corresponds to the excess pressure attaining its minimum value.

The radial separation between the successive crests at any given instant of time is λ , and hence the time interval between the arrival of the successive crests is $\frac{\lambda}{v}$, which is the time period T of oscillation at the point under consideration, where $T = \frac{1}{\nu}$, $\nu (= \frac{\omega}{2\pi})$ being the frequency of sound heard by an observer stationed at the point under consideration.

All these considerations get modified when the source of sound and the observer are in motion relative to the medium through which the wave propagates. Fig. 9-23 depicts a number of crests emitted from a stationary source (S), showing that the successive crests reach an observer (O), also stationary in the medium, *at the same rate* (ν) at which they are emitted from the source. In other words, for this special case of a stationary source and a stationary observer, one has

$$\nu' = \nu, \quad (9-42)$$

where ν' is the frequency of the sound as noted by the observer.

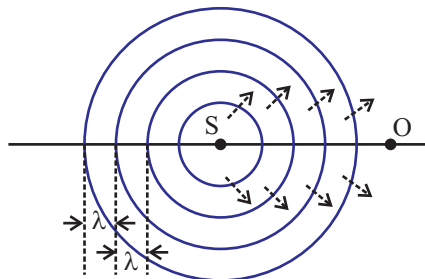


Figure 9-23: A set of successive crests emitted by a stationary source (S); these reach a stationary observer (O) at the same rate as they are emitted by the source.

Fig. 9-24, on the other hand, depicts a number of successive crests emitted by a source *moving* through the medium with a velocity, say, V where, for the time being we assume a motion with a uniform velocity along a straight line. Since the source itself proceeds through the medium as the successive crests are emitted, the disposition of the crests at any given instant of time looks as in the figure, where one observes that the separation between the successive crests in front of the source is less compared to the separation behind it. Thus, for an observer at rest in front of the source the successive crests arrive in quicker succession compared to the rate at which the crests reach an observer located behind the source. In other words, the motion of the source results in ν' , the frequency noted by the observer being *different* from ν , the frequency of emission of the crests by the source. As seen from the figure, ν' will be greater than ν for an observer in front of the source, and less than ν for an observer located behind the source.

Fig. 9-24 also shows the path of a *moving* observer (dotted line). It is evident that, due to the motion of the observer, she encounters the successive crests at a rate ν' not only different from the frequency ν of emission from the source, but one that may even vary with time since the rate at which she crosses the successive crests depends on whether, at the instant under consideration, she is located in front of or behind the source or, more precisely, on her instantaneous position relative to the train of crests.

In summary, the frequency of sound (ν') recorded by an observer differs from the frequency (ν) emitted by the source when the source and the observer are in motion relative to the medium in which the sound propagates. This phenomenon goes by the name of *Döppler effect*.

Though I have presented the basic idea underlying the explanation of Döppler effect by referring to sound waves set up in a fluid medium, the effect itself is more general in its scope, and can be observed in other instances of wave motion as well. In particular, Döppler effect is a phenomenon of considerable importance in the propagation of electromagnetic waves. In the latter case, wave propagation can take place even without any material medium, and the deviation of the observed frequency ν' from the

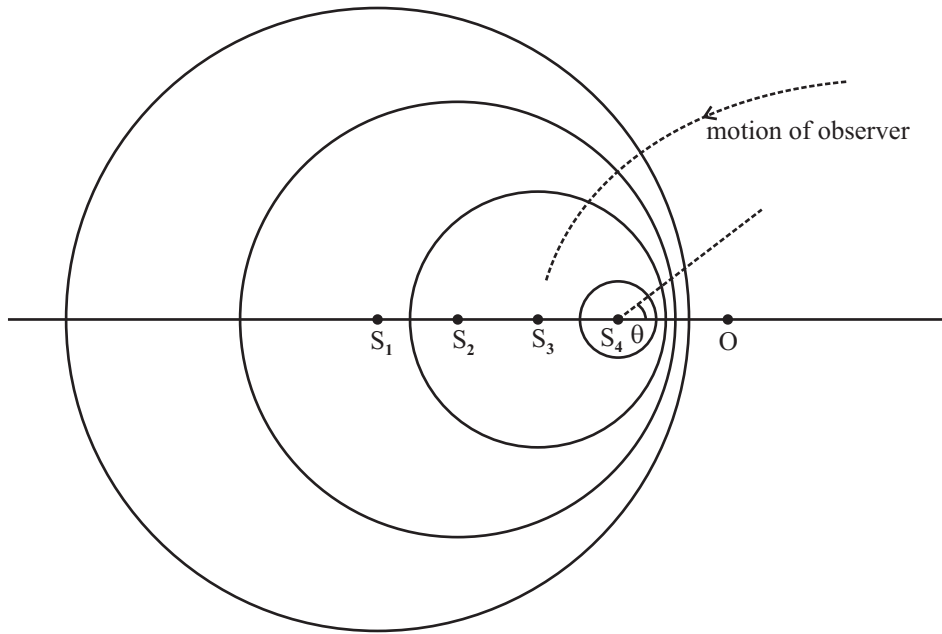


Figure 9-24: A set of successive crests emitted by a moving source, depicted at a given instant of time; S_1 , S_2 , S_3 , S_4 are four successive positions of the source; the crests are crowded in front of the source as compared to their spacing behind it; an observer (O) at rest in the medium in front of the source encounters the successive crests at a rate different from that for one located behind the source; the dotted line shows the path of a *moving* observer; the rate at which the observer crosses the successive crests now depends on the instantaneous position of the observer with respect to the train of wave fronts, and may be, in general, time-dependent; assuming that S_4 denotes the instantaneous position of the source at the instant at which the crests are located as shown in the figure, a straight line drawn at an angle θ to the line of motion of the source gives the separation between successive crests measured in that direction (see sec. 9.14).

emitted frequency ν depends on the *relative velocity* of the observer with respect to the source.

9.13.2 Frequency related to rate of change of phase

In order to make the above ideas more precise and to work out the relation between ν , the emitted frequency, and ν' , the observed frequency, it is convenient to start from the following mathematical expression relating the frequency to the phase:

$$\nu = \frac{1}{2\pi} \frac{d\Phi}{dt}. \quad (9-43)$$

This equation needs some attention since it is of considerable use in various contexts, and applies equally well to emitted and observed frequencies. Since a source of a con-

stant frequency ν emits the crests at a uniform rate $\frac{2\pi}{T}$ regardless of whether it is in motion or not (recall that successive crests which corresponds to a phase difference $\delta\Phi = 2\pi$ are emitted at a time interval T , where T stands for the time period of the source), the rate of change of phase is given by $\frac{d\Phi}{dt} = \frac{2\pi}{T}$, from which follows eq. (9-43).

In the general case in which the phase of the wave emitted by the source varies non-uniformly with time, the expression (9-43) may be adopted as the definition of the frequency of the wave. This general definition of the frequency applies, for instance to the *frequency modulated* waves used in telecommunication. In the present context, however, we assume for the sake of simplicity that the source emits waves of a constant frequency. The basic formula of Doppler effect (refer to eq. (9-51) and (9-52) below) applies for a source of variable frequency as well.

Considering next an observer in any arbitrary state of motion, if the phase of the wave at the instantaneous position of the observer be Φ at any given point of time t , then the frequency recorded by her will again be given by the formula (9-43), which simply expresses the rate at which she encounters successive crests of the wave. In other words, it applies to the observed frequency as well as to the emitted frequency where, in each case the phase Φ is to be interpreted appropriately. In the special case of the source and the observer at rest in the medium under consideration, the formula gives the same value of the frequency for the source and for the observer.

Finally, the formula (9-43) remains valid regardless of the nature of the wave (whether a spherical wave, plane wave, or a wave of some other description), provided only that it is described by a wave function of the form

$$\psi(x, y, z, t) = A(x, y, z, t) \exp(i\Phi(x, y, z, t)), \quad (9-44)$$

where the function $A(x, y, z, t)$ varies slowly with time compared to the time variation of $\Phi(x, y, z, t)$ (in which case A and Φ are referred to as the ‘amplitude’ and ‘phase’ functions describing the wave). This includes, as a special case, a wave for which the amplitude

function is time independent.

9.13.3 Döppler effect:the general formula

Let us now assume for the sake of generality that the source and the observer are both in motion in the medium, in which the velocity of sound is v . Let the instantaneous position of the source be denoted by $\mathbf{r}(t)$, while that of the observer be denoted by $\mathbf{r}'(t')$ (see fig. 9-26; see below for the definition of t' for any given t). Let us consider a wave front corresponding to the phase Φ emitted by the source at time t (when the source is at the position $\mathbf{r}(t)$), and let this same wave front reach the observer at time t' , when the observer is at $\mathbf{r}'(t')$. Thus, the wave front must have traveled through the distance $|\mathbf{r}(t) - \mathbf{r}'(t')|$ in time $t' - t$. Since the velocity of propagation of the wave front is v , we have the following equation describing the propagation of the wave front,

$$|\mathbf{r}(t) - \mathbf{r}'(t')| = v(t' - t), \quad (9-45)$$

expressing the ‘observer-time’ t' implicitly as a function of the ‘source-time’ t .

In addition, an application of the formula (9-43) to the source and to the observer in succession gives

$$\nu = \frac{d\Phi}{dt}, \quad \nu' = \frac{d\Phi}{dt'}, \quad (9-46)$$

i.e., in other words,

$$\nu'(t') = \nu(t) \frac{dt}{dt'}. \quad (9-47)$$

This is the basic formula of Doppler effect expressing the observed frequency ν' in terms of the emitted frequency ν where the derivative $\frac{dt}{dt'}$ is to be evaluated by making use of the functional dependence of t' on t determined implicitly by eq. (9-45).

9.13.4 Uniform motions of source and observer

As a simple application of the above general formula describing the shift in frequency in Doppler effect, let us assume that the source and the observer both are in uniform motion along a given line, say the x-axis, with velocities V and V' respectively. The instantaneous positions of the source and the observer are then given by

$$x(t) = x_0 + Vt, \quad x'(t) = x'_0 + V't, \quad (9-48)$$

where x_0 and x'_0 are constants, corresponding to the initial positions of the source and the observer.

Eq. (9-45) then assumes the form

$$x'_0 + V't' - x_0 - Vt = v(t' - t). \quad (9-49)$$

Starting at the point $x_0 + Vt$ at time t and moving with speed v , the wave front reaches the point $x'_0 + V't'$ at time t' .

This gives

$$\frac{dt}{dt'} = \frac{V' - v}{V - v}, \quad (9-50)$$

or, in other words,

$$\nu' = \nu \frac{v - V'}{v - V}. \quad (9-51)$$

In this formula, v represents the velocity of sound in the medium in which the source and the observer are located. If the medium itself moves (with respect to a given frame in which V and V' are defined) with a velocity whose component along the x-axis is v_0 then one has to replace v in the above equation with $v + v_0$. Alternatively, V and V' in formula (9-51) are to be interpreted as the velocities of the source and the observer along the x-axis (i.e., the common direction of their motion) relative to the medium, in which case v will denote the velocity of sound with respect to the medium, assumed to

be at rest (reason this out).

In the above formula it has been implicitly assumed that the observer lies to the right of the source (i.e., in the direction of increasing x) throughout the time interval under consideration and that V and V' carry their own signs. For instance, if the source moves along the positive direction of the x -axis and the observer along the negative direction then the sign of V will be positive while that of V' will be negative. One can generalize by saying that V and V' are the velocities of the source and the observer reckoned in the direction from the source to the observer, in which case the implied assumption is that the source and the observer do not cross over during the interval under consideration. In the event of a cross-over, the signs of each of the velocities V and V' before and after the event will get changed.

Problem 9-5

A source of sound emitting a note of frequency 2000 s^{-1} moves with uniform velocity $V = 10 \text{ m}\cdot\text{s}^{-1}$ along the y -axis of a co-ordinate system. What is the frequency of the note heard by an observer located on the x -axis at $x_0 = 40 \text{ m}$, if the time of emission from the source is 3 s from the moment the source moves past the origin (velocity of sound = $350 \text{ m}\cdot\text{s}^{-1}$?). At what time is the sound of this frequency recorded by the observer?

Answer to Problem 9-5

HINT: Considering a time t after the passage of the source through the origin O (fig. 9-25), if t' be the time at which the observer hears the sound emitted by the source, located at position P (say; not marked in the figure), at time t , one has

$$t' = t + \frac{\sqrt{V^2 t^2 + x_0^2}}{v}.$$

Thus, using eq. (9-47), one gets

$$\nu' = \frac{\nu}{1 + \frac{V}{v} \frac{Vt}{\sqrt{V^2 t^2 + x_0^2}}},$$

which is nothing but the result obtained from eq. (9-51) by putting $V' = 0$ and by replacing V with $V \cos \theta$ (check this out), the latter being the instantaneous longitudinal component (i.e., the

component along the line joining the source *at time t* and the observer) of the source velocity. In this sense, one says that there is no *transverse* Doppler effect in acoustics.

Substituting given values, one gets $\nu' = \frac{2000}{1 + \frac{10 \times 30}{350}} \text{ s}^{-1} = 1966 \text{ s}^{-1}$ (approx). The time of recording of this frequency by the source, obtained from the above relation between t and t' , is $t' = 3 + \frac{50}{350} \text{ s} = 3.143 \text{ s}$.

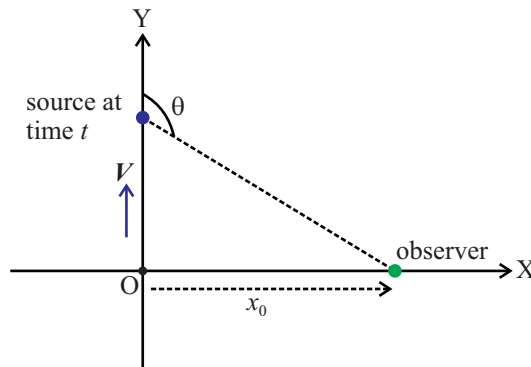


Figure 9-25: Depicting a situation in which a source moves with uniform velocity along the y-axis of a co-ordinate system, the observer being stationary at a point on the x-axis, at a distance x_0 from the origin, as in problem 9-5; θ is the angle between the direction of the source velocity and the line joining the source position (at time t , the time at which a wave front is emitted by the source, the same wave front being received by the observer at time t') with the (fixed) observer position.

Problem 9-6

A source of sound emitting a note with a frequency $\nu = 1000 \text{ s}^{-1}$ starts from a point O on the x-axis of a co-ordinate system with an initial velocity $u = 0$ and moves with a uniform acceleration $a = -5.0 \text{ m} \cdot \text{s}^{-2}$ along the x-axis (i.e., the motion occurs actually along the negative direction of the x-axis). What is the frequency of the note heard by a stationary observer located at a point with co-ordinate $x_0 = 100 \text{ m}$ at a time $t' = 10 \text{ s}$ from the time of commencement of motion of the source (velocity of sound = $350 \text{ m} \cdot \text{s}^{-1}$)?

Answer to Problem 9-6

The relation between t' , the time at which the observer records the frequency and the time (t) at which the sound is emitted by the source is given in the present case by $t' = t + \frac{1}{v}(x_0 - (ut + \frac{1}{2}at^2))$,

where v stands for the velocity of sound. Thus, using eq. (9-47) $\nu' = \nu[1 - \frac{1}{v}(u + at)]^{-1}$. This is the same as formula (9-51) with $V' = 0$, and with V replaced with $V(t)$, the velocity of the source *at the instant when it emits the sound*. What is relevant from the point of view of the observer, however, is the time t' at which she hears the note. Thus, one has to solve for t in terms of t' , which works out to

$$t = \frac{v}{a}(1 - \frac{u}{v})[1 - \sqrt{1 - \frac{2a(vt' - x_0)}{v^2(1 - \frac{u}{v})^2}}]$$

(out of the two possible roots of the quadratic equation in t , one is to be discarded to ensure that, for $t' = \frac{x_0}{v}$, t has to be 0). One thus has, in terms of t' ,

$$\nu' = \frac{\nu}{(1 - \frac{u}{v})\sqrt{1 - \frac{2a(vt' - x_0)}{v^2(1 - \frac{u}{v})^2}}}.$$

This reduces to formula (9-51) (with $V' = 0$) for $a = 0$ (uniform motion of the source, for which $V = u$). Referring to given data and putting (in SI units) $u = 0$, $a = -5.0$, $x_0 = 100$, $t' = 10$, $v = 350$, one gets $\nu' = 885\text{s}^{-1}$ (approx).

More generally, making use of eq. (9-47), one can establish the result

$$\nu'(t') = \nu(t) \frac{v - V'(t') \cos \phi}{v - V(t) \cos \theta}, \quad (9-52)$$

where v stands for the velocity of sound, and $V(t)$ and $V'(t')$ denote respectively the speed of the source at time t and the speed of the observer at time t' , and θ , ϕ are the angles shown in fig. 9-26, i.e., $V'(t') \cos \phi$ is the component of the velocity of the observer (at time t') along the line joining S (position of source at time t) and O (position of observer at time t') in fig. 9-26, and $V \cos \theta$ is similarly the component of the velocity of the source (at time t) along the line SO. Once again, if the medium in which the source and the observer are located is a moving one, then one has to replace v in the above equation with $v + v_0$, where v stands for the velocity of sound in the stationary medium, and v_0 denotes the component of the velocity of the medium along SO.

The Döppler formulae arrived at in the above paragraphs make sense only if the change in the shifted frequency (ν') in one complete time period relative to the observer (T' ; $T' = \frac{1}{\nu'}$) be small compared to the frequency ν' itself i.e., if

$$T' \frac{d\nu'}{dt'} \ll \nu', \quad (9-53)$$

which is therefore the condition under which the formulae (9-47), and (9-52) hold (in the case of (9-51), the left hand side of the above inequality is zero).

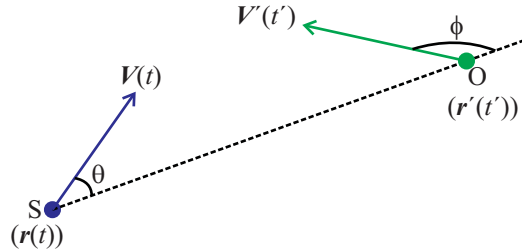


Figure 9-26: Illustrating the general situation in Döppler effect; S denotes the position of the source at time t , when it emits a wave front with a certain phase Φ ; the same wave front is received by the observer at time t' ; $\mathbf{V}(t)$ and $\mathbf{V}'(t')$ are the velocities of the source and the observer at times t and t' respectively, while $\mathbf{r}(t)$, $\mathbf{r}'(t')$ denote the corresponding position vectors with respect to some chosen origin (not shown); ϕ and θ are the angles made by $\mathbf{V}'(t')$ and $\mathbf{V}(t)$ with the line SO, i.e., with the vector $\mathbf{r}'(t') - \mathbf{r}(t)$.

Use of the Döppler effect in echo-cardiography.

In echo-cardiography, a beam of an ultrasonic wave of appropriate frequency is directed to the heart of a person and the reflected beam is recorded so as to obtain an image of the heart, to be used as an aid for diagnostic purposes. However, a simple record of the image does not furnish information regarding blood flow characteristics in the blood vessels around the heart. On the other hand, when the frequency (ν') of the reflected wave from a particular region of the heart is compared with that (ν) of the wave used as the probe, the velocity of flow in the blood vessels in that region can be determined (refer, for instance, to formula (9-51) which indicates the dependence of the frequency of the wave recorded by an observer on the velocity of the source - in the present context, the velocity of the blood cells from which the ultrasound is reflected).

small

A number of issues relating to the Döppler effect have not been taken up here though I feel that these are of sufficient interest and relevance to deserve at least a passing mention.

1. First comes the question of the *frame of reference*. It might appear at first that the result (9-51) (or, more generally, the formula (9-52)) is valid in the frame of reference in which the medium (in which the acoustic wave is set up) is at rest, the velocity of sound in this frame being v , a characteristic of the medium itself, determined by a number of its physical properties. However, while the velocity of sound in the medium is indeed of relevance in the Doppler formula, the formula itself is valid in any other frame moving uniformly with respect to the medium. One has then to interpret all the velocities appearing in (9-51) as velocities *relative* to that frame. In other words, the formula remains the same in any and every frame moving uniformly relative to the medium.

This, indeed, is as it should be, since it is required by the *principle of equivalence*. Of course, all our considerations are confined to the *non-relativistic* setting, where the velocities concerned are small compared to c , the velocity of *light* in vacuum. For velocities comparable to c , the formula will look different since terms of higher degrees in the velocities (measured in units of c) will appear, in which case it will conform to a principle of equivalence of a broader validity compared to the principle formulated in the non-relativistic setting, namely the *relativistic* principle of equivalence. In reality, however, such a formula is of little relevance since, even for much smaller velocities, different sets of phenomena make their presence felt in the generation and propagation of the elastic waves like, for instance, the *shock waves* introduced in sec. 9.14.

2. This contrasts with the Doppler effect of *light* or, more generally, of electromagnetic waves. Electromagnetic waves can be set up even in vacuum, in which they move with the velocity c *regardless* of the frame of reference (where it is to be assumed that the latter belongs to a particular class, namely the one made up of the *inertial* frames of reference). Here the universal constant c appears in the Döppler formula along with the velocities of the source and observer, i.e., in other words, the 'velocity of the

medium' relative to the frame of reference plays no role. As a result, the velocities of the source and observer appear in the formula only in the form of velocities relative to the frame of reference chosen. It goes without saying that the Doppler formula for electromagnetic waves conforms to the *relativistic* principle of equivalence (refer to chapter 17). In the case of electromagnetic waves set up in a medium, on the other hand, the velocity of the medium makes its appearance, and the formula acquires features analogous to those of the Doppler formula for acoustic waves.

3. Finally, the Doppler effect for acoustic waves, as derived in the non-relativistic setting, is wholly a *longitudinal* one, in which the frequency shift depends on the components of the velocities along the line joining the source and the observer (where the positions of the two are to be considered at two different instants, namely the time when the wave starts from the source and the time when it reaches the observer), and the transverse components do not appear in the formula. This again contrasts with the relativistic case where there is a transverse Doppler effect as well. The difference is essentially due to the fact that, in the relativistic setting, the transformation formulae from one frame of reference to another involve a *transformation of time*, in addition to the transformation of position coordinates of an event (refer, once again, to chapter 17).

9.14 Supersonic objects and shock waves

Looking at fig. 9-24, one notes that the crowding effect of the successive crests in front of a moving source as compared to those behind the source gets enhanced as the velocity of the source is made to increase. While the radial distance between the successive crests for a stationary source is λ , a simple calculation shows that, for a moving source with velocity V , the separation between successive crests directly in front of the source is $\lambda(1 - \frac{V}{v})$ and that directly behind the source is $\lambda(1 + \frac{V}{v})$, where v stands for the velocity of sound in the medium.

Considering a line inclined at an angle θ with the direction of motion of the source (see fig. 9-24), the separation between the successive crests is seen to be $\lambda(\sqrt{1 - \frac{V^2}{v^2} \sin^2 \theta} - \frac{V}{v} \cos \theta)$. This reduces to $\lambda(1 - \frac{V}{v})$ and $\lambda(1 + \frac{V}{v})$ for $\theta = 0$ and $\theta = \pi$ respectively.

However, these formulae apply only for $V < v$, i.e., for a source whose velocity is less than that of sound in the medium. For a source with $V \rightarrow v$, the separation of the successive crests in front of the source goes to *zero*: all the crests catch up with one another and get piled up in front of the source, resulting in a *high pressure front* developing just ahead of the source, the front surface being perpendicular to its direction of motion (fig. 9-27). For a source moving with a velocity $V > v$, the source cuts through the train of wave fronts, leaving a conical *wake* behind it (fig. 9-28) referred to as a bow wave by analogy with the wake left behind a high-speed motor-boat moving on water. An object moving through a medium with a velocity greater than the velocity of sound in the medium is called a *supersonic* one, its *Mach number* (i.e., the velocity measured in units of the velocity of sound ($\frac{V}{v}$)) being greater than unity.

A supersonic object, regardless of whether it is a source of sound, generates such a conical wake behind it where the tip of the cone gets flattened into a high-pressure front, namely, the *shock front* with a large energy density associated with it. Such a front, which moves through the medium under consideration, is also referred to as a *shock wave*. The pressure and energy, both of which are concentrated on the surface of the front, get dissipated in the medium, creating an intense rumbling sound, termed a *sonic boom*.

Phenomena analogous to those associated with supersonic objects and shock waves are observed in the context of wave motions of other descriptions as well. In the case of electromagnetic waves, a charged particle moving through a medium with a velocity larger than the phase velocity of light through the medium is found to radiate energy along a conical surface, a phenomenon referred to as *Cerenkov radiation*.

Fig. 9-28 depicts the train of wave fronts left behind by a supersonic object and the conical surface representing the bow wave. The angle θ made by a generator of the cone with the line of motion of the object is given by

$$\sin \theta = \frac{v}{V}, \quad (9-54)$$

where $\frac{V}{v} (> 1)$ is the Mach number of the object. I repeat that while the wave fronts shown in the figure correspond to waves of some particular frequency, such as sound waves emitted by the object under consideration, the formation of the shock front itself does not require the object to be a source of sound since it is formed by the disturbances created in the medium due to the motion of the object.

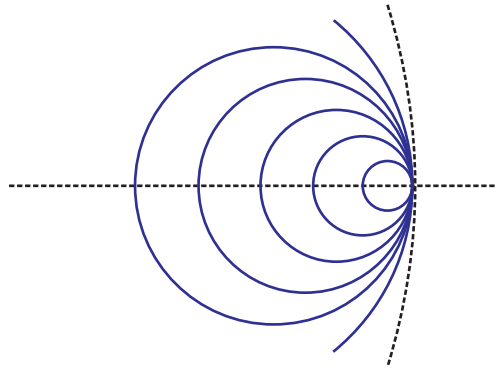


Figure 9-27: Shock front developing in front of the source as the source velocity tends to the velocity of sound ($V \rightarrow v$); the shock front is formed at right angles to the direction of motion of the moving source.

Ultrasonic shock waves are used in the medical procedure of *lithotripsy* where a shock wave is directed at one or more ‘stones’ (*calculi*) formed in the gall bladder or the kidney of a patient to break the stones into small fragments and render them relatively harmless.

As we indicate below, the shock front is part of a nonlinear wave, which is to be distinguished from the bow wave left behind a supersonic object. The term ‘shock wave’ actually refers to this nonlinear wave that leads to the formation of the shock

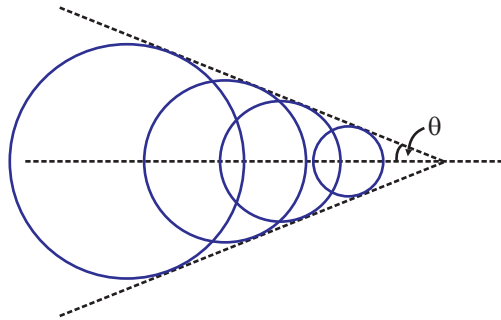


Figure 9-28: Bow wave in the wake of a supersonic source; depicting the angle θ (see formula (9-25)); the train of spherical wave fronts left behind in the wake of the source is shown.

front.

9.14.1 The production of shock fronts

Shock waves are produced and transmitted in compressible fluids, especially in gaseous media (recall that sound waves also require compressible media for propagation and transmission). More pertinently, a shock front does not always need a supersonic object for its generation. Any sudden disturbance due to which a mass of fluid is suddenly set in motion is associated with irregularities and perturbations that propagate in the form of elastic waves in the medium in front of the moving mass. Referring to the case of a supersonic object, it may be noted that the latter drags a mass of fluid along with it, as a result of which the elastic waves are set up in layers of moving fluid in front of the object, and wave fronts move ahead of the object to more distant regions where the fluid is at rest or is moving at a low velocity.

Considering a train of successive wave fronts, say a train of crests, the ones in layers close to the front of the moving fluid have a velocity greater than the ones that have moved ahead since the latter are now moving in stationary fluid. In addition, the temperature in the moving fluid is also somewhat higher than that of the fluid ahead since the fluid gets heated locally in virtue of low thermal conductivity that makes its motion nearly adiabatic, and this makes the velocity of the distant crests lower than the velocity of those just in front of the moving fluid. In summary, as the successive wave fronts move ahead, those that are at a distance from the front of the moving fluid mass

have a speed less than the ones closer to the fluid front, as a result of which the latter catch up with the more advanced wave fronts and, in course of time, there occurs a *piling up* of the wave fronts in regions ahead of the moving fluid mass.

The moving fluid mass may be one carried along by a supersonic object, or perhaps one created in an explosion of some kind, suddenly propelling the mass with a high velocity and at a high pressure. Fig. 9-29 depicts a simplified representation in which A denotes the front of a moving mass of fluid, with the flow velocity diminishing to the right of A so that the fluid at B is almost stationary. Imagine that a wave front (say, a crest) having originated at some earlier time (t_0) from the front (A) of the moving fluid is located at B at some chosen instant ($t > t_0$). B_1, B_2, B_3 are wave fronts located at the same instant (t), having been produced at earlier times t_1, t_2, t_3 subsequent to t_0 ($t_0 < t_1 < t_2 < t_3 < t$). Observe that the separations between the successive wave fronts increases as one moves back from B to A because of the speed differential between the successive wave fronts indicated above. This leads to the piling up of wave fronts alluded to earlier, due to which the wave disturbance ahead of A, originally created as a small amplitude sound wave, acquires a *large amplitude*, the propagation of which is no longer described by a linear wave equation.

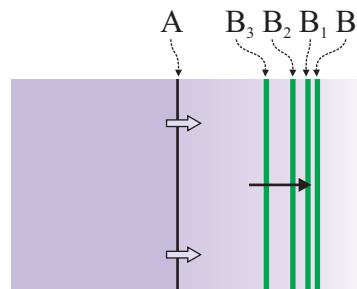


Figure 9-29: Depicting schematically the piling up of wave fronts thereby leading to the production of a large amplitude nonlinear wave as shown in fig. 9-30; A represents the front of a mass of fluid (e.g., one carried by a supersonic object) moving in from the left; disturbances in the fluid produced near A propagate ahead of the front toward the right in the form of elastic waves, but the velocity of the waves decreases downstream (see text) and there occurs a piling up of wave fronts (say, the crests); as a crest, produced at time t_0 reaches the position B at time t , crests produced at subsequent instants t_1, t_2, t_3 arrive at B_1, B_2, B_3 ; the resulting crowding of wave fronts generates a nonlinear wave and, eventually, a shock front; a one dimensional propagation is depicted for the sake of simplicity of presentation.

The large amplitude nonlinear wave propagating ahead of A is characterized by the distinctive feature that the high pressure regions of the wave profile propagate at a larger velocity compared to low pressure ones (the crests move at a higher pace compared to the troughs), as a result of which the profile deviates more and more from a sinusoidal one with a self-sustained process of steepening of the profile as depicted in fig. 9-30, till there develops a region in the profile where the wave function (i.e., the pressure or velocity) tends to become multi-valued. At this stage the wave profile starts falling apart and *turbulence* sets in, with *irreversible processes* beginning to play a dominant role in the wave dynamics on a cascade of smaller and smaller length scales (these processes arise due to effects of viscosity and thermal conduction). This causes the region of multi-valuedness to be replaced with a steep and thin front within which *entropy production* takes place because of the irreversible dissipative processes. In contrast, the flow remains almost reversible and adiabatic (i.e., *isentropic*) on either side of the front. This, in bare outlines, is how a shock front is generated.

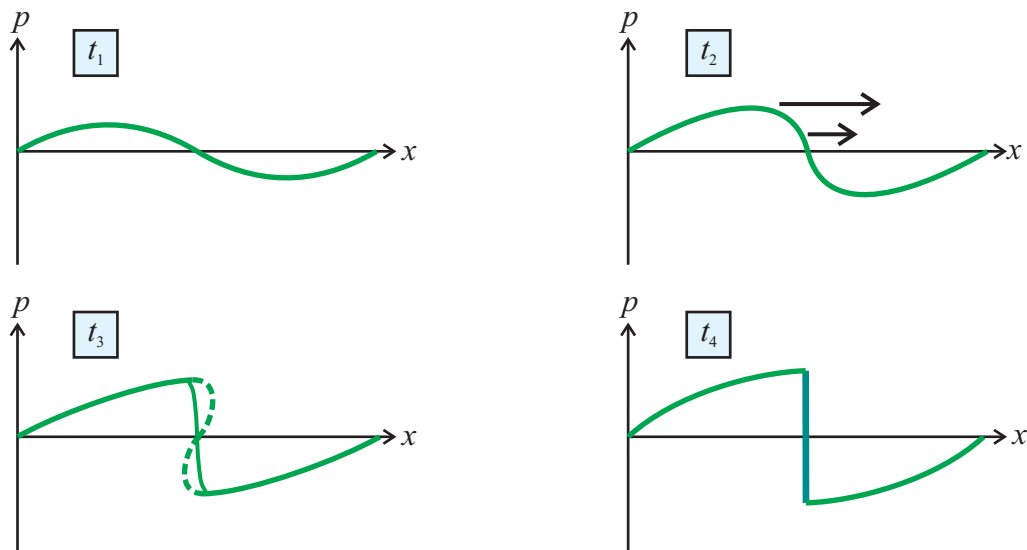


Figure 9-30: The process of self-steepening of a nonlinear wave with progressively increasing amplitude; the pressure (p) is plotted against distance of propagation (x) at four successive time instants ($t_1 < t_2 < t_3, t_4$); at time t_1 the profile is sinusoidal because of a small amplitude; at time t_2 the amplitude has increased and the wave profile deviates from a sinusoidal one since the high pressure regions tend to move ahead of the relatively low pressure ones; the resulting steepening of the wave profile is a self-reinforcing process and at time t_3 , there arises a thin region in the wave profile where the pressure tends to be multi-valued (dotted line); this leads to an instability where dissipative processes acquire overriding importance in a thin layer of the fluid, generating a shock front (time t_4).

Fig. 9-31 depicts the variation of pressure on either side of a shock front located at some position P , with the moving mass of fluid (not shown) located to the left. The fluid to the right of P (i.e., downstream) is almost at rest and at a low pressure, since the shock is yet to pass through this region. The fluid upstream, on the other hand, moves in with a large velocity and at a high pressure. The shock front, in other words, represents a *surface of discontinuity* of pressure and velocity. Generally speaking, the velocity of the shock wave is higher than the speed of sound in the stationary fluid, but less than the velocity of the fluid mass moving in from the high pressure side of the shock.

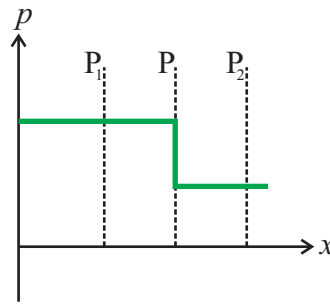


Figure 9-31: Depicting the pressure discontinuity across a shock front; the pressure is plotted against the distance along the fluid in the direction of propagation of the front; at some specified instant of time, P denotes the location of the front while P_1 , P_2 are locations upstream and downstream respectively (the fluid mass moving in from the left is not shown); the front propagates towards the right; an element of fluid at P_2 approaches the front, crosses it, and then occupies a position behind it; in the process, its pressure increases and specific volume decreases in an irreversible process as depicted in in fig. 9-32; the fluid velocity also decreases discontinuously across the front.

Referring to fig. 9-31, we observe that as the shock front moves past some location in the fluid (say, the one at P), a small mass of fluid located at a point such as P_2 ahead of it at some given point of time moves in the opposite direction relative to the front and thus eventually gets transferred to a location behind the front. This flow through the front causes an *irreversible* increase in the pressure (p) and specific volume (v) of the fluid - one that cannot be depicted in the p - v diagram by a continuous path. In contrast, the passage of a sound wave corresponds to a reversible motion of the representative point (one representing an equilibrium state in the p - v diagram) along an adiabatic curve. Fig 9-32 depicts such a reversible path followed by the representative point from A to B along an adiabatic curve corresponding to the passage of a sound wave where a pressure

minimum (trough) of the wave profile moves past some given location in space, being succeeded by a crest. This is to be compared with the passage of a shock front through the same location where the representative point now jumps from A to C, the dotted line in the p - v diagram connecting the two points indicating that an irreversible process has taken place during the passage of the shock front. The relation between p_A, v_A and p_C, v_C can be derived from conditions of conservation of mass, momentum and energy of the fluid element under consideration as it crosses the front (the entropy of the element, however, increases) and the resulting formula - of great practical value - is referred to as the Rankine-Hugoniot relation, which involves additional thermodynamic parameters. As mentioned above, the irreversible process from A to C is an adiabatic one, though not isentropic.

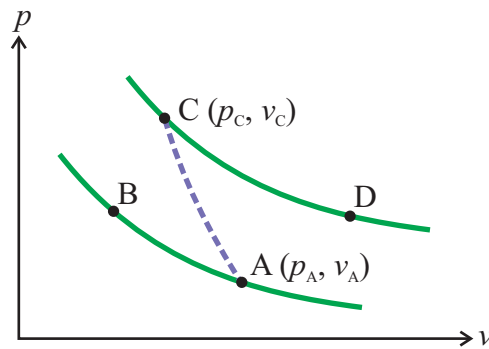


Figure 9-32: Depicting the irreversible change in pressure (p) and specific volume (v) of an element of fluid as it crosses a shock front; AB and CD are adiabatic curves in the p - v diagram; in the case of the passage through the wave front of a sound wave the representative point at A moves to B along the adiabatic curve and then moves back to A, as the pressure and specific volume of the element of fluid under consideration oscillates sinusoidally; in contrast, the passage through a shock front leads to an irreversible change in which the representative point moves to C (dotted line), but the process cannot be represented by a continuous path in the p - v diagram; while mass, momentum, and energy is conserved in the process, entropy increases in virtue of dissipation involving viscosity and thermal conduction; the relation between p_A, v_A and p_C, v_C is referred to as the Rankine-Hugoniot formula.

9.15 Superposition effects

The principle of superposition was introduced in sec. 9.5.3. Several interesting phenomena in wave motion can be explained in terms of this principle. In the context of

acoustics, the principle of superposition applies to small amplitude waves, which are described by wave equations of the form (9-15).

9.15.1 Interference

9.15.1.1 Introduction to the idea of coherence

Interference of waves will be discussed at considerable length in the context of wave optics (see sec. 15.3). In the present section, the basic idea underlying the phenomenon of interference will be outlined in the context of acoustic waves.

Fig. 9-33 depicts two point sources of sound (S_1, S_2) from which acoustic waves of the same frequency and wavelength reach the point of observation O. We assume that the waves are represented by expressions of the form (see eq. (9-16))

$$p_1 = \frac{A}{r_1} \cos(kr_1 - \omega t), \quad p_2 = \frac{A}{r_2} \cos(kr_2 - \omega t), \quad (9-55)$$

where r_1 and r_2 stand for distances from the two sources and the constant A is assumed to be the same for the two waves (corresponding to identical power output from the two sources; more generally, two different constants, say A_1, A_2 , could be assumed to characterize the waves without fundamentally altering the conclusions in this section). In addition, the waves from the two sources have been assumed to be *coherent* where, once again, the idea of coherence will be discussed at greater length in the context of electromagnetic waves and wave optics (see sections 14.10, and 15.3). In the present context, the concept of coherence can be briefly explained as follows.

Recall from sec. 9.4 that the expression for excess pressure may contain a term such as δ in the phase part, termed the initial phase or epoch.

The initial phase was introduced in the context of the one dimensional plane wave in sec. 9.4. However, essentially the same idea applies for spherical waves as well.

Thus, one could introduce, for the sake of generality, initial phase terms, say δ_1, δ_2 in

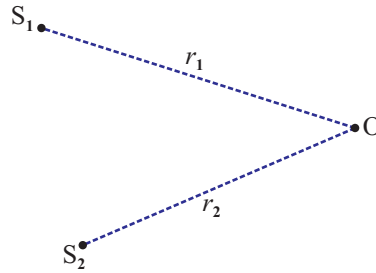


Figure 9-33: Illustrating the idea of interference; S_1 and S_2 are two coherent sources of sound (say, a pair of speakers powered by the same amplifier) sending out spherical waves to the point of observation O where the two waves interfere; the intensity at O depends on its position relative to the sources, and there results a spatial pattern consisting of maxima and minima of intensities; this interference pattern is determined by the *phase relations* between the two waves at the various observation points; for the observation point O , the intensity is determined by the path difference between the two ray paths, i.e., by $r_1 - r_2$.

the expressions for p_1 , p_2 in eq. (9-55). However, while talking of effects involving just the two waves without reference to any other wave, one can take one of the initial phases, say, δ_1 to be zero. This leaves us with the other phase δ_2 , which we now call δ for the sake of convenience where, to be precise, δ stands for the *difference* of the initial phases of the two waves.

The two waves are said to be *coherent* with respect to each other, when they are characterized by the same wavelength and frequency and when, moreover, the constants A (more generally, A_1 , A_2), δ have well-defined constant values (we have assumed δ to be zero for the sake of simplicity while writing the expressions for p_1 , p_2 in eq. (9-55)). There may be situations, on the other hand, where these constants (of which δ is of special importance) effectively behave as *random* variables, when the waves under consideration are said to be mutually *incoherent*.

1. The above definition of coherence can be expressed by saying that coherence is the consequence of a high degree of *correlation* between the two waves under consideration. A lack of correlation between the waves finds its mathematical expression in the constants A and δ behaving effectively like random variables.
2. As an example of coherent acoustic waves, imagine two loudspeakers fed from the same microphone and powered by the same amplifier. The waves emitted by the loudspeakers will then be mutually coherent. If, on the other hand, the

loudspeakers are powered by two different amplifiers then the waves emitted from these will involve uncorrelated noisy variations, and will be incoherent.

9.15.1.2 Interference as superposition of coherent waves

Referring back to fig. 9-33 and the expressions (9-55) for the waves from the two sources S_1, S_2 , which we have assumed to be coherent, the resultant wave function at any point like O is given by $p = p_1 + p_2$, in accordance with the principle of superposition. As the position of the point O is made to vary, the resultant wave function changes since r_1 and r_2 , the distances of O from the source points S_1 and S_2 get changed. This variation of r_1 and r_2 affects the wave function p through the amplitude terms ($\frac{A}{r_1}$ and $\frac{A}{r_2}$) as also through the phases $\Phi_1 = kr_1 - \omega t$ and $\Phi_2 = kr_2 - \omega t$. Typically, it is the variation of the phases that affects the wave function to a greater degree than does the variation of the amplitudes.

Thus, for the sake of simplicity, we may altogether ignore the variation of the amplitude of either wave with the change in position of the point O, and represent the two waves in the form

$$p_1 = A \cos \Phi_1 = A \cos(kr_1 - \omega t), \quad p_2 = A \cos \Phi_2 = A \cos(kr_2 - \omega t). \quad (9-56)$$

Making use of these simplified expressions, we can have an idea as to how the wave function $p = p_1 + p_2$ for the superposed wave varies as the observation point O is made to shift from one location to another. Assume, for instance, that the observation point O is such that $r_1 = r_2$, and thus, $\Phi_1 = \Phi_2$. The wave functions p_1 and p_2 then get added up to an amplitude $2A$, and the resultant intensity at O is four times the intensity due to either of the two waves (recall from sec. 9.11.4 that the intensity is proportional to the square of the amplitude; while this rule was derived in the context of a plane wave, it is of more general validity and holds for spherical waves as well).

The same value of the resultant intensity is obtained for observation points for which the difference between the phases Φ_1 and Φ_2 is any integral multiple of 2π . The two

waves are said to interfere *constructively* at such points.

Suppose now that the point O is shifted to a new position for which $\Phi_1 - \Phi_2 = \pi$ or, more generally, an odd integral multiple of π . The resultant wave function and intensity at O is then *zero*. The two waves are said to interfere *destructively* at such points.

Thus the superposition of the two coherent waves under consideration results in a spatial variation of intensity: the intensity varies from point to point, being a maximum at some points, namely, those where the waves interfere constructively, and a minimum at some others, the latter being the ones where the waves interfere destructively. If, instead of the superposed wave, one considers the spatial variation of intensity due to either of the two waves, then one finds a monotonic variation (almost a constant over a considerable region of space) instead of this alteration of intensity between maxima and minima.

Moreover, for the superposed wave, the variation of intensity with the location of the point of observation in space *differs from a simple sum of intensities due to the waves from the two sources considered independently of each other*. In other words, the superposition of the coherent waves gives rise to a *redistribution* of the energy flux in space, depending on the variation of phase difference between the two waves. It is this phenomenon of redistribution of energy flux that goes by the name of *interference*.

The spatial pattern involving maxima and minima of intensity generated through the interference of two or more waves is referred to as an interference pattern, which is said to consist of *interference fringes*, a fringe being typically a locus of constant (usually a maximum or a minimum) intensity. While I have explained the basic idea underlying the explanation of interference by referring to a pair of coherent spherical waves, essentially the same explanation applies to plane monochromatic waves as well.

An acoustic wave generated from a source can be a coherent (or *pure*) one or else, may be an incoherent superposition, or a *mixture*, of an number of coherent components. Depending on the way the various components are mixed up, the wave under consideration

may be characterized by a certain *degree* of coherence. For instance, the plane progressive wave produced by a source which is only approximately a monochromatic one, can be expressed in terms of its *coherence length*, a concept explained in section 15.3.6.4.

The condition for two waves to generate an interference pattern is that the coherence length of the wave emitted by the source has to be larger than the path difference between the two ($|r_1 - r_2|$ in the above example) for all points of observation in the region in which the pattern is to be formed. Because of the comparatively large wavelength of acoustic waves, these are typically characterized by a large coherence length. This is why, interference of sound waves may be observed over a wider range of situations as compared to light waves since, in the case of sound waves, the restrictions imposed by the condition of coherence are less severe.

Incidentally, from the expressions of the phases in (9-56), one infers that the relation between the path difference ($\Delta = r_1 - r_2$) and the phase difference ($\delta\Phi = \Phi_1 - \Phi_2 = k(r_1 - r_2)$) for the two waves reaching any given point at a given instant of time is

$$\delta\Phi = \frac{2\pi}{\lambda} \Delta. \quad (9-57)$$

Interference patterns involving light waves are formed in a *Young's double slit pattern*, and in *Newton's rings* in optics (see chapter 15).

Problem 9-7

Two sources of sound S_1 and S_2 (see fig. 9-33) emit monochromatic waves of the same frequency f and at the same phase, the distance between the two sources being 0.3m. A recording device is kept at O (fig. 9-33) at a distance of 0.4m from S_2 , where the line joining S_2 and O is perpendicular to that joining S_1 and S_2 . At what frequencies within the range 30s^{-1} and $30,000\text{s}^{-1}$ will O record a minimum intensity (velocity of sound, $v = 350\text{m}\cdot\text{s}^{-1}$)?

Answer to Problem 9-7

HINT: From the given geometry of the problem, the path difference between the waves reaching O from the two sources is $\Delta = 0.1$ m (check this out) which has to be $(2n + 1)\frac{\lambda}{2}$ for destructive interference to occur ($n = 0, 1, 2, \dots$; this corresponds to (see (9-57)) $\Phi_1 - \Phi_2 = (2n + 1)\pi$ in the notation introduced above). Since $f = \frac{v}{\lambda}$, the required condition reduces to $f = (n + \frac{1}{2})\frac{v}{\Delta} = (n + \frac{1}{2}) \times 3500\text{s}^{-1}$. This lies within the given frequency range for $n = 0, 1, 2, \dots, 8$.

9.15.2 Standing waves

9.15.2.1 Standing waves in an air column

Fig. 9-34 shows a tube filled with air, where both the ends of the tube are closed. It is fitted with a rod entering through one end, with a diaphragm attached to the end of the rod inside the tube. The rod can be set in vibration by drawing a piece of leather along its length, which sets up a vibration of the diaphragm and the latter, in turn, causes an oscillation to take place in the layers of air in the tube, where the excess pressure at each point executes a simple harmonic oscillation. The variations of the excess pressure in contiguous layers of air take place in a correlated manner because of the stress force exerted by one layer on an adjacent one. The resulting space-time dependence of the excess pressure is once again described by a wave equation of the form (9-14), where v stands for the velocity of sound in air.

However, now the solution does not look like a plane progressive wave of the type (9-6) because of the fact that the wave has been set up in a tube with closed ends, and the excess pressure has to satisfy a set of *boundary conditions* at these ends. The latter can be stated as the requirement that the displacement of the air particles at each of the ends has to be zero. The resulting excess pressure distribution within the tube constitutes a *stationary* (or *standing*) wave.

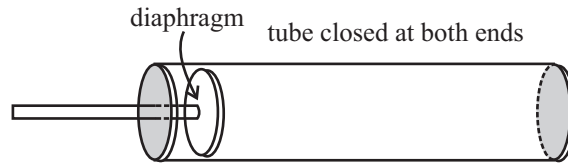


Figure 9-34: Set-up for generating a stationary wave in a tube closed at both ends.

9.15.2.2 Features of a standing wave

For a sufficiently narrow and long tube the variation of excess pressure turns out to be given by an expression of the form

$$p(x, t) = p_0 \cos(kx) \cos(\omega t), \quad (9-58)$$

where ω is a characteristic frequency that can assume any one of a *discrete* set of values depending on the length of the tube, and k is likewise a constant having any one of a discrete set of values (being related to ω as $k = \frac{\omega}{v}$). In the above expression, x stands for the co-ordinate of any point in the tube measured along a line (the x -axis) along the axis of the tube, with the origin being chosen at one end (say, the left end in the figure), and p_0 is a constant, referred to as the amplitude of the stationary wave.

One can check by direct substitution that the expression (9-58) does indeed satisfy the one dimensional wave equation (9-14) (check this out). Let us now see how it relates to the boundary conditions at the ends of the tube.

To start with, we note that the way the stationary wave is set up by the arrangement shown in fig. 9-34 looks similar to that in fig. 9-20 where the wave was similarly assumed to be set up by the vibrations of a diaphragm, with the difference, however, that the vibrations of the successive layers of air (or of the fluid medium in which the wave is set up) did not have to obey the boundary conditions applying at the two ends of the tube since the medium was assumed to be unbounded on one side of the vibrating diaphragm.

The waves set up in an unbounded medium also have to satisfy a boundary condition,

namely, the *boundary condition at infinity* which I have left implied in the discussion of sec. 9.11.

The relations between the excess pressure, longitudinal strain, and displacement indicated in sec 9.11.3 (see eq. (9-23)) hold in the present context as well, and the expression for the displacement s consistent with eq. (9-58) is seen to be

$$s = -\frac{p_0}{kK} \sin(kx) \cos(\omega t), \quad (9-59)$$

where the boundary conditions to be satisfied by the wave are given by

$$s = 0 \text{ at } x = 0, x = L \text{ for all } t, \quad (9-60)$$

L being the length of the tube. While the boundary condition at $x = 0$ is automatically satisfied by the expression (9-59) (indeed, the expression 9-58 has been chosen so as to conform to this boundary condition), the boundary condition at $x = L$ requires

$$k = \frac{n\pi}{L}, \quad (n = 1, 2, \dots). \quad (9-61a)$$

Corresponding to these possible values of the constant k characterizing the standing wave solutions (of the form (9-58)) in the tube, the possible values of ω , the frequency characterizing the standing wave are

$$\omega = \frac{n\pi}{L}v = \frac{n\pi}{L}\sqrt{\frac{\gamma P}{\rho}}, \quad (n = 1, 2, \dots). \quad (9-61b)$$

where the meanings of the symbols have already been explained.

These possible frequencies of the standing wave are integral multiples of

$$\omega_0 = \frac{\pi v}{L}, \quad (9-62)$$

referred to as the *fundamental frequency* for the tube under consideration.

The time dependent part in either of the expressions (9-58) and (9-59) can be replaced, for the sake of generality, with the more general form $\cos(\omega t + \delta)$ where the constant δ represents an initial phase. I have chosen an initial condition for which $\delta = 0$ so as to keep things simple while looking at the principal features of standing waves.

9.15.2.3 Superposition of propagating waves

Before discussing at greater length the principal characteristics of standing waves expressed by the above equations, I have to tell you what all this has to do with the principle of superposition.

While the monochromatic propagating and standing waves introduced in this chapter are solutions of the wave equation (eq. (9-14) or (9-15) as the case may be) subject to appropriate boundary conditions and while, in the ultimate analysis, it is the wave equation that provides the definition and description of these waves (or of any small amplitude acoustic wave, for that matter), there may exist alternate ways, specific to the situations under consideration, of describing and analyzing the waves.

In the case of the standing waves in the tube, for instance, one may describe a standing wave as the wave resulting from the *superposition* of two one dimensional plane progressive waves propagating in opposite directions along the tube. In other words, one may adopt the following convenient picture describing the setting up of a standing wave: a monochromatic plane wave propagating along the x-direction gets reflected from the end face of the tube at $x = L$, thereby producing a similar plane wave propagating along the *negative direction* of the x-axis (towards the left in fig. 9-34). The two waves, one propagating towards the right and the other towards the left, then get superposed with each other, producing the standing wave.

Indeed, the expression (9-58) results from the superposition of the wave functions given by

$$p_1 = \frac{p_0}{2} \cos(kx - \omega t), \text{ and } p_2 = \frac{p_0}{2} \cos(kx + \omega t), \quad (9-63)$$

these being the two plane progressive waves referred to above. This tells us that a standing wave can indeed be looked upon as a superposition of two oppositely directed progressive waves.

This way of looking at a standing wave as a superposition of two propagating waves is more a convenient picture describing the standing wave than a description of the actual mechanism by which such a wave is set up. Thus, even a progressive wave, in turn, may be imagined to result from a superposition of two standing waves.

9.15.2.4 Progressive waves and standing waves: a few points of distinction

Despite the fact that the expression (9-58) results from a superposition of the expressions in (9-63), a standing wave differs fundamentally from a propagating wave. For one thing, a standing wave described by, say, eq. (9-58), represents a simple harmonic vibration at every point x with a frequency ω , where all these vibrations for various different values of x *are in the same phase, though with varying amplitudes*. For instance, the amplitude of vibration at the point x is given by $p_0 \cos(kx)$ while the phase is given by $\Phi = \omega t$ for all x . In the case of a progressive wave, on the other hand, the vibrations occur with the same amplitude at all points (p_0 for the wave represented by eq. (9-4)), but *with differing phases* since the expression for the phase is $\Phi = kx - \omega t$.

More precisely, a standing wave is made up of successive segments, where all points in any given segment vibrate in phase with $\Phi = \cos \omega t$, while the points in the next segment vibrate in the *opposite* phase, with $\Phi = -\cos \omega t$. Thus, a segment between two successive *nodes* (see below; refer to sec. 9.16.1 where standing waves resulting from transverse vibrations of a stretched string are introduced and to fig. 9-39 where the formation of segments in a standing wave is illustrated) vibrates in the opposite phase compared to the next segment between two successive nodes.

Another fundamental point of distinction between the two relates to the fact that the possible values of the frequency of a standing wave form a *discrete* set (eq. (9-61b);

correspondingly, the possible values of k also form a discrete set given by (9-61a)), while the possible frequencies (and corresponding values of k) form a continuously distributed set of values (ranging, in principle, from 0 to ∞) for a propagating wave.

Finally, a *standing wave does not involve an energy flux* while a propagating wave certainly does. Thus, the energy with which the vibrations in the pipe in fig. 9-34 is set up, cannot flow out of the tube and gradually gets dissipated in the air and in the surrounding material bodies. On the other hand, in the set-up described by fig. 9-20 the energy spent in driving the vibrations of the diaphragm flows out into the medium through the propagating wave set up in it. The intensity at any point in the medium provides a measure of the rate of flow of energy per unit area of a surface imagined at that point, oriented perpendicularly to the direction of flow of the energy.

9.15.2.5 Modes of standing waves in an air column

Looking at equation (9-58) along with (9-61a), (9-61b), one can say that a standing wave can be set up in a tube in any one of a number of possible *modes* (also referred to as *normal modes* of vibration of the air column in the tube), where each mode is characterized by an integer n ($n = 1, 2, \dots$). The value $n = 1$ is said to correspond to the *fundamental* mode of vibration with frequency ω_0 given by eq. (9-62). The frequencies of the other possible modes are all integral multiples of this fundamental frequency, and are referred to as its *harmonics*, where the frequency of the n th harmonic ($n = 2, 3, \dots$) is $n\omega_0$.

Corresponding to each possible value of ω , there occurs a value of k given by $k = \frac{\omega}{v} = \frac{2\pi}{\lambda}$, where λ is *twice* the distance between two successive points in the standing wave having the same amplitude (by contrast, in the case of a progressive wave, $\lambda = \frac{2\pi}{k}$ gives the distance between two successive points with the same *phase*). For instance, for the mode characterized by the integer n , the points given by

$$x_m = \frac{L}{n}m, \quad (m = 0, 1, 2, \dots, n), \quad (9-64)$$

correspond to a maximum value of the amplitude of the excess pressure, namely $|p_0|$ (a vibration of the form $p_0 \cos(\omega t)$ and one of the form $-p_0 \cos(\omega t)$ can be said to be characterized by the same amplitude $|p_0|$, but with opposite phases; unless otherwise stated, we choose $p_0 > 0$ in eq. (9-58)). These points are referred to as *pressure antinodes*. On the other hand, the points given by

$$x_m = \frac{L}{n}(m + \frac{1}{2}), \quad (m = 0, 1, 2, \dots, n - 1), \quad (9-65)$$

correspond to *zero* amplitude of the excess pressure, and are referred to as *pressure nodes*.

One observes that the distance between two successive pressure nodes or that between two successive pressure antinodes is given by $\frac{L}{n}$ which, as mentioned above, is half the value of $\lambda = \frac{2\pi}{k} = \frac{2L}{n}$ for the n th mode under consideration.

We have seen that the vibrations at two consecutive pressure antinodes occur with opposite phases. If, on the other hand, we consider two consecutive antinodes with the *same* phase, then the separation between the two equals the value of $\lambda = \frac{2\pi}{k}$ for the mode under consideration.

Analogous to the pressure nodes and the pressure antinodes considered above (where one refers to eq. (9-58)) one may consider displacement nodes and displacement antinodes by referring to eq. (9-59). It transpires from such a consideration that *displacement nodes coincide with pressure antinodes* and, similarly, displacements *antinodes* with pressure *nodes* (check this out).

While the normal modes considered above describe special standing wave solutions of the one dimensional wave equation subject to the boundary conditions appropriate to the tube closed at both ends, more general standing wave solution can be produced in the tube by the superposition of more than one normal modes. A standing wave obtained by such a superposition does not correspond to a simple harmonic time variation of the wave function (say, the excess pressure) at any given point, nor does it have a

simple structure involving nodes and antinodes. But the superposition still represents a standing wave since the energy flux associated with such a wave continues to be zero.

Standing waves in one dimension may also be formed with boundary conditions other than the ones corresponding to a tube with both ends closed. For instance, standing waves can be set up in a long and narrow tube with one end closed and the other end open, or in one with both ends open. Such a set-up will then be characterized by its own set of allowed wavelengths (analogous to those implied by formula (9-61a)) and of normal mode frequencies.

Standing waves in three dimensions

The standing waves formed in long narrow tubes are all solutions to the one dimensional wave equation, subject to appropriate boundary conditions. Three dimensional standing waves are also possible in chambers or closed cavities of various regular shapes. These are, at times, referred to as *resonators*. A resonator is characterized by a fundamental frequency, as also by the frequencies of the higher harmonics, and is capable of producing a loud distinctive sound when an object, say a tuning fork, vibrating at any of these frequencies is brought in front of it. The vibrations of the standing waves in the cavity (typically, the fundamental mode in it) resonate with those of the body brought near the resonator because of a matching of the two frequencies involved (see sec. 4.6). Indeed resonance is a commonly invoked way of reinforcing or amplifying the sound produced by a source. Thus, a tuning fork with its stem pressed against a hollow wooden box produces a sound louder than that produced by the tuning fork vibrating on its own because the vibrations of the tuning fork set up a resonance with some mode or other of the standing waves produced in the air-filled cavity of the box.

Acoustic standing waves play an important role in the musical sounds produced in flutes and a number of other musical instruments like the pipe organ. In addition, the resonating effect produced by an object vibrating in front of an air cavity is made use of in various other musical instruments.

Standing waves are formed in various other contexts as well. *Optical standing waves* are of great importance in physics. Standing waves built up in carefully constructed optical resonators are of relevance in the production of intense and coherent *laser* light. Bound stationary states of electrons in an atom can be described as standing waves where the wave function of an electron (see chapter 16 for an introduction of a few basic ideas in this context) satisfies an equation of a similar nature as the wave equation (9-15) subject to an appropriate boundary condition.

9.15.3 Beats

In sections 9.15.1 and 9.15.2 we looked at superposition effects involving waves of identical frequencies. The superposition of waves with slightly *different frequencies* produces the phenomenon of beats. Consider, for instance, two waves given by the expressions

$$p_1 = p_0 \cos(kx - \omega t), \quad p_2 = p_0 \cos(k'x - \omega' t), \quad (9-66)$$

where, for the sake of convenience, the waves are assumed to be propagating in the same direction (along the x-axis) and where

$$\omega' = \omega + \delta\omega, \quad k' = \frac{\omega'}{v} = k + \frac{\delta\omega}{v} = k + \delta k \text{ (say)}, \quad (9-67)$$

$\delta\omega$ ($\ll \omega$) being a small difference in the two angular frequencies. In the expressions (9-66), the amplitudes of the two waves are assumed to be the same, also for the sake of simplicity.

The superposition of the two waves produces the resultant

$$p = p_1 + p_2 = 2p_0 \cos\left(\left(\frac{\delta k}{2}\right)x - \left(\frac{\delta\omega}{2}\right)t\right) \cos\left(\left(\frac{k+k'}{2}\right)x - \left(\frac{\omega+\omega'}{2}\right)t\right). \quad (9-68)$$

In this expression the first of the two cosine terms varies *slowly in space and time* since δk , $\delta\omega$ are small quantities. Indeed, with $\delta\omega \ll \omega$, and consequently, $\delta k \ll k$, an appreciable variation of this term with x occurs only over distances large compared to the wavelength of either of the two waves, and similarly, an appreciable variation in time

occurs only over intervals large compared to the time period of either wave. Therefore, considering relatively smaller regions of space and smaller intervals of time, the first cosine term can be assumed to be effectively a constant while the second cosine term is seen to represent a progressive wave with frequency $\frac{\omega+\omega'}{4\pi}$ and wavelength $\frac{4\pi}{k+k'}$. The time-dependent amplitude at any given point is thus

$$A(x, t) = 2p_0 \cos\left(\left(\frac{\delta k}{2}\right)x - \left(\frac{\delta\omega}{2}\right)t\right). \quad (9-69)$$

The intensity at the point under consideration, which is proportional to the square of the amplitude is then of the form

$$I = I_0(1 + \cos((\delta k)x - (\delta\omega)t)), \quad (9-70)$$

where I_0 is an appropriate constant. For any given x , the cosine term varies with time t in the range -1 to $+1$, and thus the intensity varies slowly and periodically (with a time period $T_{\text{beat}} = \frac{2\pi}{\delta\omega}$) between 0 and $2I_0$. The frequency of variation of the intensity is

$$\nu_{\text{beat}} = \frac{\delta\omega}{2\pi} = \nu_1 - \nu_2, \quad (9-71)$$

where $\nu_1 (= \frac{\omega_1}{2\pi})$ and $\nu_2 (= \frac{\omega_2}{2\pi})$ are the frequencies characterizing the two waves under consideration (we assume $\nu_1 > \nu_2$ without loss of generality).

In other words *the intensity due to the superposition of the two waves with nearly equal frequencies varies periodically with a frequency given by the difference of the two frequencies*. The slow variation of intensity of the sound resulting from the superposition is referred to as 'beat'.

Beats are also produced by the superposition of plane waves propagating in two different directions, as also by the superposition of two spherical waves or even two waves with more complex spatial structures, the condition of beat formation being that the time dependence of each of the two waves is to be a simple harmonic one, the two frequencies being almost equal.

Problem 9-8

A tuning fork, held at some distance in front of an wooden barrier, emits a note of frequency $\nu = 510\text{s}^{-1}$, while the barrier is made to move towards the tuning fork at a speed of $5\text{m}\cdot\text{s}^{-1}$. If the velocity of sound is $350\text{m}\cdot\text{s}^{-1}$, find the frequency of the beats formed.

Answer to Problem 9-8

HINT: Along with the acoustic wave produced by the tuning fork, one has to consider the wave reflected from the barrier, which can be interpreted as being produced by the 'image' of the tuning fork. The 'image' moves at a speed of $V = 10\text{m}\cdot\text{s}^{-1}$ towards the observer (reason this out; recall that the object distance equals the image distance at any particular instant of time), and the two waves, on being superposed, produce beats. The frequency (ν') of the reflected wave differs from ν because of the motion of the image-source, and is obtained by putting $\nu = 510$, $V = 10$, $V' = 0$, $v = 350$ (all in SI units) in formula (9-51), which gives $\nu' = \frac{35}{34}\nu$. As a result of the superposition of the two waves, beats are produced, with frequency (refer to formula (9-71)) $\delta\nu = \nu' - \nu = \frac{\nu}{34} = 15\text{s}^{-1}$.

9.15.4 Wave packets: group velocity

Looking at either of the two waves in eq. (9-66), the wave profile, i.e., the plot of, say, $p = p_1(x, t)$ against x for any chosen value of t is a sinusoidal one as shown in fig. 9-35(A), similar to the one shown in fig. 9-3 where one observes that the profile moves in space with the velocity $v = \frac{\omega}{k}$. Fig. 9-35(B), on the other hand shows the profile of the superposed wave given by eq. (9-68) (again for some chosen value of t), where only a part of the profile is shown, in which the heights of successive crests is seen to diminish on either side of a central peak. As indicated by dotted curves, similar structures recur on either side of the one shown in the figure.

Since the phase velocities of both the individual waves making up the superposed wave are the same ($v = \frac{\omega}{k} = \frac{\omega'}{k'}$, both being given by (9-25)), the entire profile shown in fig. 9-35(B) propagates with the same velocity.

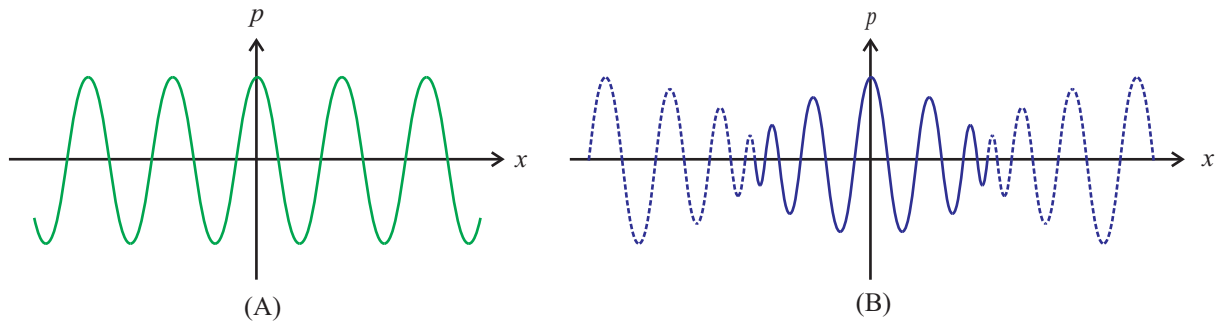


Figure 9-35: Wave profile for (A) a single plane monochromatic wave (refer to fig. 9-3), and (B) the superposed wave given by eq. (9-68); only a finite part of the profile is shown in (B), where the heights of the successive crests is seen to diminish on either side of a central peak; similar structures recur on either side of the one shown in the figure (indicated with dotted curves).

More generally, one may consider the superposition of an infinite number of waves with frequencies varying continuously over a range, say, ω to $\omega + \delta\omega$ (and correspondingly, with k varying over a range, say, from k to $k + \delta k$). The resulting profile (fig. 9-36) looks similar to the one shown in fig. 9-35(B) but contains just one single structure with crests of diminishing heights on either side of a central peak, there being no build-up to other similar structures on either side.

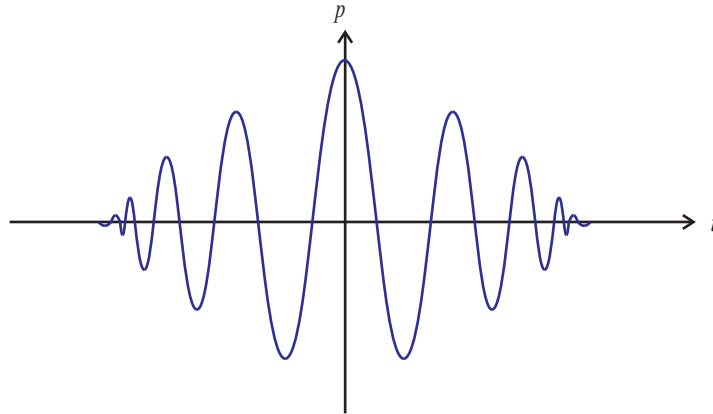


Figure 9-36: Wave packet in one dimension, resulting from a superposition of waves with frequencies spread over a range.

Such a wave profile, made up of a large number of waves superposed with one another, is termed a *wave packet*. In the case of small amplitude acoustic waves in a fluid,

each of the component waves propagates with the same velocity ($v = \sqrt{\frac{K}{\rho}}$, eq. (9-25)), in consequence to which the entire wave packet also travels with the velocity v . The fact the ratio $\frac{\omega}{k}$ of a component wave does not depend on the frequency or the wavelength of the wave, is expressed by saying that the fluid is a *non-dispersive* medium with respect to these waves.

In other instances of wave motion, however, such as electromagnetic waves in a medium (see sections 14.7, 14.9), large amplitude waves in a fluid, seismic waves, or ripples on the surface of a water body (refer to sec. 7.6.8.6), the phase velocity of waves does depend on the wavelength, a phenomenon known as *dispersion*. For such waves, the individual components making up a wave packet propagate with velocities differing from one another, as a result of which the motion of the wave packet presents interesting features.

As the individual components propagate, the packet gets altered in shape wherein it becomes, in general, more and more elongated. At the same time, it retains, to some degree, its identity as a profile with a structure like the one shown in fig. 9-36, since the structure as a whole moves with a certain velocity, referred to as the *group velocity* of the wave packet.

In this section I have considered, for the sake of simplicity, the formation of wave packets with reference to waves in one dimension. More generally, wave packets are solutions to the three dimensional wave equation formed by the superposition of plane waves with their frequencies varying over some range, and the wave vectors k also varying, in *direction* as well as in magnitude. They are called wave *packets* since their wave profiles are *localized* in space over some region or other.

A plane wave characterized by a single frequency and a single wave vector k is an idealization. Instead, wave packets formed by the superposition of plane waves with frequencies and wave vectors varying over small ranges are more commonly realized in practice. The region of space over which such a wave packet is localized may be large enough to be assumed to be effectively infinite in extent so that a wave packet may, under certain

circumstances, be pictured as a plane wave.

'Pure' plane waves (those characterized by sharply defined frequencies and wave vectors) can be 'mixed' with one another in two essentially distinct ways : by *coherent superposition*, or by mixing them while they remain *uncorrelated* with one another. The former results in a wave packet where all the component waves have definite phase correlations with one another. The latter, on the other hand, produces incoherent waves. While wave packets produce interference effects, with alternating maxima and minima of intensity in some given region of space, incoherent waves do not produce interference effects.

The distinction between coherent superposition and incoherent mixing holds not only for plane waves but for spherical waves and waves with more general spatial dependence as well.

9.16 Vibrations of strings and diaphragms

Acoustic standing waves (i.e., those generated by pressure variations in fluids) have been introduced in sec. 9.15.2. However, as mentioned there, the formation of standing waves is a wave phenomenon of a more general nature, where standing waves are found to be solutions to the wave equation subject to certain boundary conditions. In the present section we will encounter two other instances of standing waves, namely, those in the transverse vibrations of stretched strings, and of stretched diaphragms.

Analogous to a vibrating tuning fork setting up waves in the air surrounding it that cause the sensation of hearing in a listener located near the tuning fork, a stretched string or a diaphragm also sets up waves around it, thereby acting as a source of sound of any of a possible set of frequencies specific to it. The sound created by a stretched string or diaphragm can, moreover, be of a composite type involving a number of these frequencies. In other words, such a string or a diaphragm can act as a source of a rich repertoire of sounds. Vibrating strings and diaphragms are therefore of great importance as sources of sound in *musical instruments*.

9.16.1 Transverse vibrations of stretched strings

Consider a string stretched along its length and fixed at two ends as in fig. 9-37. If the string is pulled on one side of its mean position and then let go, the tension in the string produces a restoring force tending to bring the string from a configuration shown by the broken dotted line in the figure back to its mean position (straight solid line), wherein there ensues a vibration of the string, analogous to the vibrations in the air column in a long narrow tube. While the vibrations in the air column involves oscillations in the excess pressure at various points in the column, the vibrations of the string consist of transverse oscillations (i.e., oscillations in a direction perpendicular to its length) of the various points on the string.

Looking at any small element of length of the string at any time during the vibration as in fig. 9-38, the portions of the string on either side of it (indicated by dotted lines in the figure) exert forces on it due to the tension in the string. Each of the two forces acting at the end-points of the element under consideration can be resolved into two components, one along the length of the string and the other perpendicular to the length. Assuming the displacement from its mean position to be small, the components parallel to the length of the string get canceled, while the components perpendicular to the length provide a net restoring force on the element.

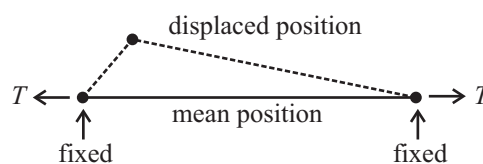


Figure 9-37: A stretched string fixed at two ends; the dotted line shows a configuration where the string is pulled on one side from its mean position; if it is now let go, a vibration of the string ensues, made up of a superposition of normal modes of standing waves.

Working out in this way the restoring force on the element under consideration, one can set up its equation of motion relating its acceleration in a direction perpendicular to the length of the string and the net restoring force. This equation is found to be of the

following form

$$\frac{\partial^2 u}{\partial x^2} = \sqrt{\frac{m}{T}} \frac{\partial^2 u}{\partial t^2}. \quad (9-72)$$

In this equation $u(x, t)$ stands for the displacement, along the direction perpendicular to the length of the string, of a point on it located at a distance x from one end (say, the left end in fig. 9-37), at any given time t , and can be taken to represent the wave function in the present context. Moreover, the symbols T and m stand for the tension in the string and its mass per unit length respectively. One observes that this equation is of the general form of a one dimensional wave equation (eq. (9-14)) where the parameter v is given by

$$v = \sqrt{\frac{T}{m}}. \quad (9-73)$$

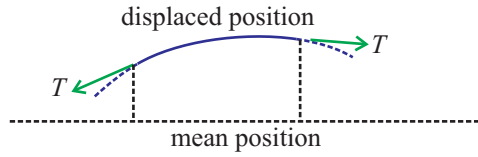


Figure 9-38: Forces acting on a small element of length (shown magnified for the sake of clarity) of the vibrating string; these forces are exerted by the portions of the string on either side of the element (indicated by dotted lines) due to the tension T in it; each of the forces acting at the two ends of the element can be resolved into two components, of which the components along the length of the string get canceled, and the components perpendicular to the length add up, providing the net restoring force for the element under consideration.

If the stretched string were infinite in length instead of being one of finite length with fixed end points, one dimensional monochromatic progressive waves, or waves resulting from superpositions of these, would be set up in it with a velocity given by the expression (9-73). For the finite string, however, a *stationary wave* will instead be set up because of the fact that the vibrations of the string will have to obey the boundary condition

$$u = 0 \text{ at } x = 0, L \text{ for all } t, \quad (9-74)$$

at the two end-points, L being the length of the string.

In the case of a monochromatic wave propagating down an infinitely long stretched string, there occurs a transfer of energy from any small segment of the string to a contiguous one. Considering any cross section of the string, the rate of flow of energy through that cross-section along the direction of wave propagation is given by the expression

$$W = \frac{1}{2}mv\omega^2 u_0^2 \quad (9-75)$$

where the symbols m and v have been defined above and u_0 is the amplitude of the transverse displacement at any point on it (we assume that the string is homogeneous and that there does not occur any dissipation of energy anywhere on the string). This may be compared with the expression (9-36) for the rate of flow of energy per unit area in the case of a propagating plane monochromatic pressure wave set up in a fluid. In both the expressions, the rate of flow of energy is proportional to the squared amplitude of the wave.

In the steady state, the energy of any given segment of the string (or of any given volume element of the fluid) remains constant since as much energy flows into the segment in any given time interval as flows out of it.

In the case of a stationary wave set up in a string of finite length, however, there does not occur any energy transfer between contiguous segments

The conditions (9-74) are analogous to the boundary conditions for the vibrations of the air column in a tube closed at both ends (eq. (9-60)) and imply that standing waves corresponding to various different *normal modes* (or, simply, *modes*) can be set up in the string. Each mode is characterized by an integer n , ($n = 1, 2, \dots$), where the wave function $u(x, t)$ for the n th mode is given by

$$u(x, t) = u_0 \sin\left(\frac{n\pi}{L}x\right) \cos\left(\frac{n\pi v}{L}t\right), \quad (9-76)$$

this being similar to eq. (9-59). In the above expression, u_0 is a constant representing

the amplitude of the standing wave under consideration at the antinodes (see below), and v is given by the expression (9-73). One observes that the variation of u at each point on the string (corresponding to any specified value of x) is a simple harmonic one with an angular frequency ω given by

$$\omega = \frac{n\pi v}{L} = \frac{n\pi}{L} \sqrt{\frac{T}{m}}, \quad (9-77)$$

In eq. (9-76), an initial phase δ can be introduced in the time dependent term for the sake of generality, which I have chosen to be zero so as to keep things simple. The constant δ is determined by the *initial conditions* of the string.

The value of k for the n th mode is given by $k = \frac{2\pi}{\lambda} = \frac{n\pi}{L}$ where λ equals twice the distance between two consecutive points on the string, characterized by the same amplitude.

Two such points on the string oscillate with *opposite phases*, i.e., with a phase difference π . If we consider two consecutive points on the string having the same amplitude *and phase*, then the distance between them equals the constant λ for the mode under consideration.

In particular, the points oscillating with the maximum amplitude ($|u_0|$; u_0 is commonly assumed to be positive, in which case an oscillation of the form $-u_0 \cos(\omega t)$ corresponds to an amplitude u_0 , but with a phase opposite to an oscillation represented by $u_0 \cos(\omega t)$) are the *antinodes* and those with amplitude 0 the *nodes*, as in the case of standing waves set up in the tube. The positions of the nodes and the antinodes for the n th mode ($n = 1, 2, \dots$), are given by

$$(\text{nodes}) \quad x_m = \frac{L}{n} m \quad (m = 0, 1, \dots, n), \quad (9-78a)$$

$$(\text{antinodes}) \quad x_m = \frac{L}{n} \left(m + \frac{1}{2}\right) \quad (m = 0, 1, \dots, n-1). \quad (9-78b)$$

The portion of the string between two consecutive nodes or between two consecutive antinodes constitutes a segment where all the points in a segment oscillate with the same phase. However, the phase of one segment differs from the next one by π , i.e., points in the two segments oscillate with opposite phases. The distance between two consecutive nodes or antinodes equals $\frac{\lambda}{2}$, i.e., $\frac{\pi}{k}$, where the value of k for any given mode is related to the frequency ω of that mode as $k = \frac{\omega}{v}$.

Fig. 9-39(A) and (B) show the two lowest modes of vibration ($n = 1, 2$) of a stretched string where it is seen that in the fundamental mode ($n = 1$) there is just one segment of the string with all the points oscillating in the same phase. In this fundamental mode the two end-points are nodes and the mid-point of the string is an antinode. In the next mode ($n = 2$), there are nodes at the two end-points as also at the mid-point of the string, while the antinodes are located at distances $\frac{L}{4}$, $\frac{3L}{4}$ from either end. The length of the string is made up of two segments in this mode where the end-points of each segment are marked by nodes. Points in these two segments oscillate in opposite phase as depicted in fig. 9-39(B).

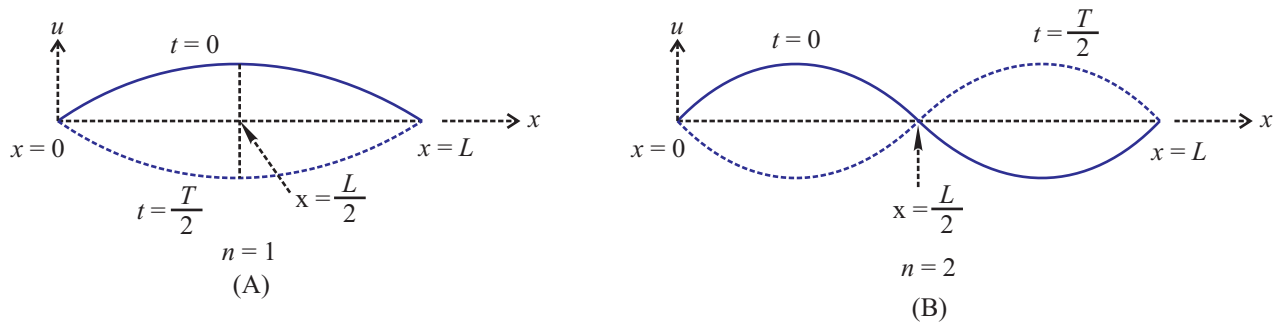


Figure 9-39: Depicting the two lowest modes of vibration of a stretched string; (A) the fundamental mode ($n = 1$); the string consists of a single segment, all points of which oscillate in phase; the solid and the dotted lines denote the profile of the string at two instants of time, at an interval $\frac{T}{2} = \frac{\pi}{\omega}$, where $\omega = \frac{\pi v}{L}$; (B) in the next higher mode ($n = 2$), the string vibrates in two segments (or loops) which are in opposite phase; there is now a node at the two ends and at the mid-point of the string, while two antinodes appear at $x = \frac{L}{4}$ and $x = \frac{3L}{4}$; once again, the solid and the dotted lines depict the profiles at an interval of time $\frac{T}{2}$.

In general, the vibrations of a stretched string can be described as a *superposition* of the normal modes, involving all the frequencies of the form (9-77). The modes with $n > 1$

are referred to as *harmonics* of the fundamental mode ($n = 1$). The frequencies of the harmonics being integral multiples of the fundamental frequency, the sound produced by a vibrating string is often a pleasing one. By contrast, a superposition of waves with frequencies not rationally related to one another generally produces a discordant sound.

The sound emitted by a vibrating stretched string in the fundamental mode ($n = 1$) resembles the sound from a *monopole* source while the next higher mode ($n = 2$) emits sound similar to a *dipole* source.

A stretched string can be set into motion either by plucking at any appropriate point through a small distance or by delivering an impulsive force by striking at some point of the string. The vibrations of a plucked string differs from those of a struck string in the proportions in which the different modes are superposed in the vibrations.

One has to make a number of assumptions in order to arrive at the wave equation (eq. (9-72)) for the vibrating string. For instance, the string has to be an *inextensible* one and the effect of external forces like the force of gravity has to be negligible on its motion. Moreover, the wave function $u(x, t)$ (i.e., the displacement at any given point and at any instant of time) has to be sufficiently small so that nonlinear terms may not be significant in the description of the motion.

Problem 9-9

A homogeneous string of mass $M = 0.01$ kg and length $l = 1.44$ m is stretched between two fixed points with a tension $T = 100$ N, and is made to vibrate in a standing wave mode made up of three segments (compare with fig. 9-39(B), where the string vibrates in two segments). Calculate the wavelength (λ) and frequency (ν) of the standing wave, and the speed (v) of traveling wave that would be set up in the string if it were of infinite length (two such waves of wavelength λ each can be imagined to be superposed in producing the standing wave in the string when the latter is stretched between fixed points).

Answer to Problem 9-9

HINT: Since the length of each segment (between two successive nodes) is $\frac{\lambda}{2}$, the required wavelength is $\lambda = \frac{2l}{3} = 0.96$ m (the specified mode corresponds to the second harmonic in the string, i.e., to $n = 3$ in eq. (9-76), (9-77)). The speed of transverse waves that can be set up in the string stretched under the tension T is $v = \sqrt{\frac{T}{m}} = 120 \text{ m}\cdot\text{s}^{-1}$ ($m = \frac{M}{l}$). The frequency of the standing wave is then $\nu = \frac{v}{\lambda} (= \frac{3}{2l} \sqrt{\frac{T}{m}}) = 125 \text{ s}^{-1}$.

Problem 9-10

An air-filled tube closed at both ends is set into its fundamental mode of vibration of the air column by resonance with a stretched string placed alongside it and vibrating in its first harmonic. If the lengths of the tube and of the string are $l_1 = 1.4$ m and $l_2 = 1.0$ m respectively, and if the mass per unit length of the stretched wire be $m = 0.01 \text{ kg}\cdot\text{m}^{-1}$, find the tension (T) in the string (velocity of sound in air, $v = 350 \text{ m}\cdot\text{s}^{-1}$).

Answer to Problem 9-10

HINT: Because of resonance between the stretched string and the air column in the tube, their frequencies are to be the same. Referring to the vibration of the air column, its wavelength in the fundamental mode is given by $\lambda_1 = 2l_1$ (nodes at both ends), and hence its frequency is $\nu = \frac{v}{\lambda_1} = \frac{v}{2l_1}$. This has to be equal to the frequency of the stretched string in its first harmonic ($n = 2$ in eq. (9-76), (9-77)), i.e., $\frac{v}{2l_1} = \frac{1}{l_2} \sqrt{\frac{T}{m}}$. Using given values, $T = 156.3 \text{ N}$ (approx).

9.16.2 Vibrations of stretched diaphragms

Fig. 9-40(A) shows a stretched circular diaphragm fixed along its boundary. If the equilibrium position of the diaphragm is disturbed by, say, striking it at the centre towards one side and then letting it go, the diaphragm vibrates and, under appropriate conditions, emits sound (as in the case of a drum). The displacement of any one small area element on the diaphragm is transmitted to neighbouring area elements by virtue of the tension in it and thus the vibration spreads throughout the diaphragm in the form of a *wave*.

The boundary condition corresponding to the circular periphery of the diaphragm being fixed, results in the wave being a *stationary* one.

The equation of motion of the diaphragm or of a stretched membrane turns to be of the form

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \frac{\rho}{T} \frac{\partial^2 u}{\partial t^2}. \quad (9-79)$$

In this equation, the equilibrium position of the diaphragm is assumed to be in the x-y plane, while $u(x, y; t)$ denotes the displacement of a point on the diaphragm located at the position (x, y) at time t . T stands for the tension in the diaphragm and ρ for its mass per unit area.

1. The tension in the stretched diaphragm is defined by imagining any small line element in it and considering the force exerted by the portion of the diaphragm (say A) on one side of the line, on the portion (say, B) lying on the other side. This force acts in a direction perpendicular to the line element and away from the region occupied by the portion B, and its magnitude per unit length of the line element is the tension T . Evidently, this definition is reminiscent of the definition of the *surface tension* of a liquid since a liquid surface behaves in a manner analogous to a stretched membrane.
2. The displacement u is assumed to be small so that non-linear terms may not be significant in the equation (9-79) describing the vibrations of the diaphragm. Moreover, the effects of external forces like that of gravity are assumed to be small so that the tension remains the same throughout the diaphragm.

Analogous to equations (9-14) and (9-15), eq. (9-79) is referred to as the *two dimensional wave equation*. While fig. 9-40(A) depicts a circular diaphragm with a fixed boundary, one may consider diaphragms of other shapes as well, where the shape of the diaphragm determines the boundary condition to be satisfied by the function $u(x, y; t)$, which is the wave function in this case. For the sake of concreteness, I refer to the circular diaphragm below, and will briefly refer to waves in diaphragms of other possible shapes at the end.

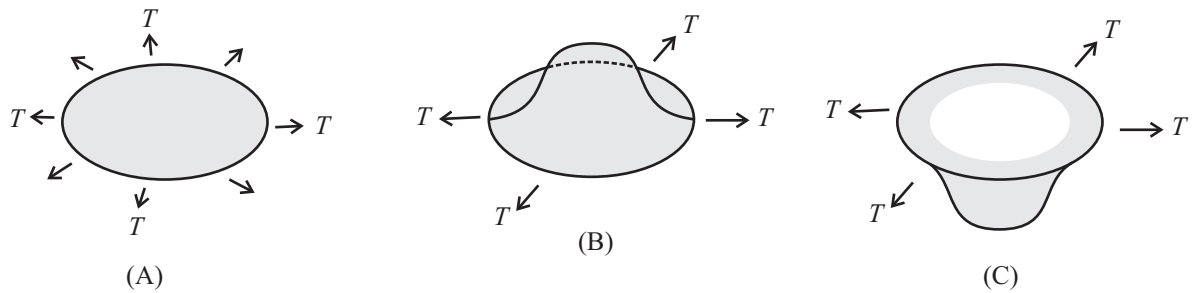


Figure 9-40: (A) A stretched circular diaphragm fixed at the boundary, under a constant tension T ; (B) and (C) depict schematically the (0,1) mode of vibration at two time instants at an interval of $\frac{T}{2}$; in (B) the displacement at the center (an antinode) is maximum in one direction while in (C), the displacement is maximum in the other direction.

The circular diaphragm can vibrate in any one of a discrete set of *normal modes* (or, simply, *modes*), where each mode is characterized by *two* integers, say, m and n . The integer m stands for the number of diametric nodal lines and n for the number of concentric circular nodal lines, a nodal line being a line on the diaphragm along which the wave function u is zero for all values of t . Incidentally, all nodal lines for the circular diaphragm are either diametric or circular lines concentric with the boundary of the diaphragm.

Vibrations of a metallic plate can be described in terms analogous to those of a membrane or a diaphragm.

Thus, in the (0,1) mode (which is the *fundamental* mode of vibration of the diaphragm) the only nodal line is its circular boundary, there being no diametric nodal line. This mode of vibration is depicted in fig. 9-40(B), (C), where the profile of the diaphragm is shown at two time instants at an interval of $\frac{\tau}{2}$, τ being the time period of the fundamental mode. A diaphragm in this fundamental mode emits sound similar to that from a monopole source. However, in this mode, the diaphragm radiates energy in the form of acoustic waves at a rapid rate, and the sound is short-lived. The diaphragm can be made to vibrate in this mode by delivering a sharp blow at the centre.

In the (1,1) and the (2,1) modes, on the other hand, there are one and two diametric

nodal lines respectively, along with the circular nodal line along the fixed boundary, the two diametric nodal lines for the $(2, 1)$ mode being perpendicular to each other. A diaphragm vibrating in the $(1, 1)$ mode acts somewhat like a dipole source of sound while one in the $(2, 1)$ mode acts like a lateral (or transverse) quadrupole source. Moreover, the vibrations in these two modes persist for a longer time compared to that in the fundamental mode, and the sound can be heard for a longer time.

A circular membrane or a diaphragm can seldom be excited in a pure normal mode of vibration, i.e., one with a definite value of m and of n . In general, the vibrations of a circular diaphragm can be described as a superposition of more than one normal modes. The relative proportions of the various normal modes in the superposition depends on the way the diaphragm (say, of a drum) is set into vibration and determines the musical quality of the sound emitted by it.

The shape of the fixed boundary of the stretched diaphragm determines its possible modes of vibration through the boundary condition imposed on its motion. For a boundary of a regular shape, like the circular or a rectangular one, the nodal and antinodal lines (an antinodal line, as opposed to a nodal line, is one where the wave function oscillates with a maximum amplitude compared to its neighboring points) are of a relatively simple structure, and the frequencies of the various normal modes can be expressed in relatively simple terms.

For most other shapes of the diaphragm boundary like, for instance, a stadium shaped one, the nodal and antinodal lines are of an extremely complex shape and the frequencies of vibration do not form any simple sequence, so that they can only be described in statistical terms. When such a diaphragm is set in vibration, the wave set up in it is, in general, of a complex type, similar to the reverberating acoustic wave set up in a closed room with an oblong or a stadium-shaped boundary.

9.16.3 Musical instruments

Musical instruments can be of various types such as stringed instruments (the violin, the guitar, the Indian *sitar*), wood-wind and brass-wind instruments (flute, clarinet, trumpet), and percussion instruments (tympani, the Indian *tabla*).

A stringed instrument, for instance, consists of a number of stretched wires of appropriate length, diameter, and mass such that when one or more of these wires are set in vibration, they emit a pleasing musical sound involving numerous *tones* of definite frequencies, all of which bear some characteristic numerical relation to one another. The musician can make delicate adjustments in these frequencies by controlling the effective lengths of the wires. The various stringed instruments differ in the manner in which the wires are set in vibration and in the ways the frequencies are controlled.

A percussion instrument, on the other hand, involves the vibrations of a diaphragm or a plate which similarly emits sound that can be described as a superposition of monochromatic components, where each component corresponds to a normal mode of the diaphragm.

In a number of these stringed and percussion instruments the sound emitted by the strings or the diaphragm is reinforced by the vibrations set up in a large hollow wooden chamber in which standing waves can be set up in a large number of normal modes. The vibrations of the strings or of the diaphragm resonates with some of these modes, producing the sensation of a rich musical sound.

In a wood-wind or a brass-wind instrument, standing waves are set up in one or more pipes whose frequencies are determined by appropriately controlling the boundary conditions and by using one or more reeds (thin strips made of appropriate material) to channelize the wind blown into the instrument. In brass-wind instruments the *mouth-piece* acts like a resonator that makes possible the selection of a range of frequencies without lengthening the main blow-pipe. In addition, the vibrations in the brass parts of the instrument enrich the sound.

The details of construction and the working of a musical instrument involves a large number of factors delicately related to one another so that many of these instruments can only be described as acoustic marvels.

9.17 Loudness, pitch, and quality of sound

The sounds that we commonly hear can be classified as *musical sound* and *noise* where the two are distinguished by the degree of regularity of the waveform depicting the excess pressure as a function of time (see fig. 9-41). A musical sound is, in the main, made up of a coherent superposition of monochromatic waves whose frequencies are related to one another in simple numerical ratios. A noise, on the other hand, is principally an incoherent mixture of waves whose frequencies do not bear any simple ratio to one another.

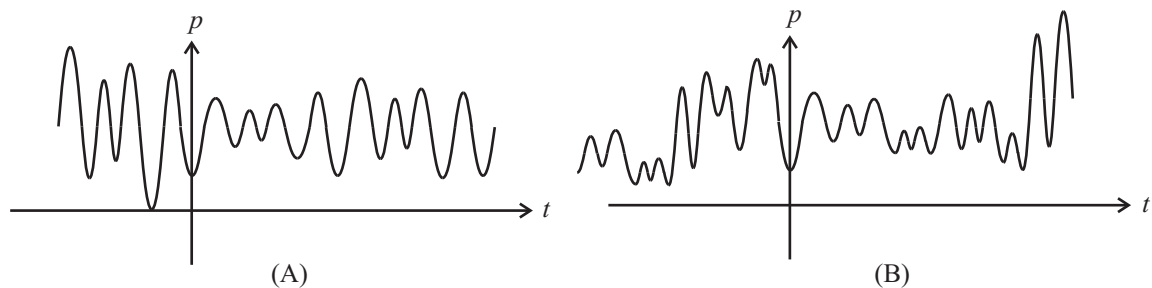


Figure 9-41: Comparing the waveforms (schematic) of (A) musical sound and (B) noise; the wave-form in (A) is smooth and regular compared to that in (B), where there is an irregular and uncorrelated variation of the wave function.

The principal ingredients of a musical sound are *notes* where a note can be described as a *unit* of sound of a certain minimum duration (commonly of a certain definite frequency), producing a certain distinct and identifiable sensation in the ear. Various different notes, or more generally, musical sounds can be described in terms of the following three *characteristic features*: the *pitch*, the *loudness*, and the *quality*.

The pitch of a note corresponds to the sensation of how shrill or how flat the sound is, and is determined mainly by the frequency of the most preponderant component in

it. However, there is no strict one-to-one correspondence between frequency and pitch, since the pitch depends to some extent on the intensity of sound as well. Thus, while the frequency is an objective characteristic of a pure note, the pitch is a related, subjective characteristic.

The loudness corresponds to the sensation of how strong or feeble the sound appears, and depends on the rate of energy flowing into the ear of the listener. It is determined by the intensity of sound produced by the source at a point close to the ear, which in turn depends on the amplitude of vibration of the source of sound and its distance from the listener. The loudness of a note also depends to some extent on the frequency (or combination of frequencies) of the sound.

The quality (or timbre) of a musical sound is that characteristic which distinguishes it from another sound of the same pitch and loudness, and is principally determined by the proportion of the various components of different frequencies making up the sound under consideration. Even a 'pure' note (sound of a definite frequency) is often mixed with small proportions of overtones (i.e., sounds of frequencies that are integral multiples of the basic frequency of the note), whose presence gives the sound a distinctive quality.

In addition, another important characteristic of a musical note is its *duration*, i.e., the time during which it persists in the ear.

9.18 Elastic waves in solids

In the present chapter, I have mainly been concerned with pressure waves in fluids while, at the same time, indicating a number of features of waves in general. The sensation of sound is most commonly generated by pressure waves in air, though sound may be produced in other fluids as well such as in underwater hearing.

Elastic waves may be produced in solid media as well. For instance, the standing waves set up in a vibrating metal plate or the vibrations of a tuning fork are the result of

variations of stress and strain at the various different points in the plate or the tuning fork. Such waves set up in a solid may, in turn, produce acoustic waves in the air surrounding it, as in the case of the tuning fork.

The elastic properties of solids are not as simple as those of fluids since a fluid is characterized by only one single elastic coefficient while an isotropic solid is characterized by two such coefficients (see chapter 6).

9.18.1 Vibrations in a crystalline medium: normal modes

Solids are mostly of crystalline nature, where the atoms and molecules are arranged in regular spatial structures. The equilibrium configuration of such a crystalline structure is built up by a repetitive arrangement of a basic *unit cell* in which a number of atoms are placed at fixed positions, both the unit cell and the positions of the atoms in it being characteristics of the material in question.

When an atom or a group of atoms in such a crystalline structure is displaced from the equilibrium position, restoring forces come into play whereby these atoms execute oscillatory motions. The inter-atomic forces in the solid then cause a transmission of these oscillations to more distant atoms, as a result of which *waves* are set up in the crystalline structure. Waves set up in an infinitely extended crystalline medium are progressive in nature while those in a finite medium are of the nature of standing waves.

These waves in a crystalline solid are, in general, of a complex nature since these are, in reality, waves involving the strain and stress fields set up in the medium, where the strain or stress at any point is described by a *tensor* rather than a scalar or vector (by contrast, recall that the stress at any point of a fluid is effectively a scalar in nature). Depending on the nature of symmetry of the crystal structure, there may be several elastic coefficients characterizing a crystalline solid, where the *anisotropy* of the structure plays a significant role. Correspondingly, the elastic waves in the solid may also be of various different descriptions. These elastic waves may once again be described as superposition of normal modes, but a complete description of all the possible normal modes is, in general, a problem of considerable complexity.

If the wavelength characterizing a normal mode is *large* compared to the average separation between the atoms in the crystal, the discreteness of the crystalline structure can be ignored so that, for all such waves, the solid can be assumed to be a *continuous medium*. Moreover, a solid body of macroscopic dimensions is often made up of a large number of small crystalline pieces, and is effectively an *isotropic* one. As I have mentioned in sec. 6.6.6, such an isotropic solid is characterized by *two* independent elastic constants, which one can take as, say, the Young's modulus (Y) and the shear modulus (η). Consequently, the description of elastic waves in isotropic solids is a relatively simple matter compared to those in an anisotropic solid. Elastic waves of large wavelengths set up in a solid are commonly referred to as *sound* waves.

In summary, a complete description of elastic waves in a solid should take into account its crystalline structure, including the discreteness of the atoms arranged in the crystal lattice and the possible anisotropy of the structure. In practice, however, most solid materials are effectively isotropic and, moreover, elastic waves of all but the smallest wavelengths (i.e., those comparable to the average separation between the atoms in the crystal) can be described by assuming the material to be a continuous medium.

9.18.2 Elastic waves in an isotropic solid

In an isotropic solid that can be assumed to be a continuous medium, *two* independent wave motions are possible as compared to only one type of wave that can be set up in a fluid, namely, a pressure wave. For instance, considering a plane monochromatic wave with a wave vector \mathbf{k} and a frequency ω , two waves can be set up in the solid (assumed to be of infinite extent) with those values of \mathbf{k} and ω .

Recall that the waves we are considering here are ones with long wavelengths, for which the solid can be assumed to be a continuous medium.

In one of these two waves, the displacement \mathbf{u} of the solid particle at any given point (more precisely, the displacement of a small volume element of the solid, where the displacements due to thermal motions of the molecules have been averaged away) is

found to be along the direction of \mathbf{k} , while in the other, it is perpendicular to \mathbf{k} . Thus, one of the two elastic waves with a given ω and \mathbf{k} is a *longitudinal* one while the other is a *transverse* wave. The transverse wave can be described as a *shear* wave since it does not involve any volume strain in the material. The longitudinal wave, however, involves longitudinal and bulk strains developed in the material due to the propagation of the wave. Finally, it is found that the velocity of the longitudinal wave is always larger than that of the transverse wave.

1. Waves in an isotropic solid can be described as *vector* waves in as much as the displacement vector \mathbf{u} at any point acts as the wave function describing these waves. By contrast, the pressure waves in fluids are *scalar* waves since the wave function (the excess pressure) is a scalar (vector wave functions can be defined in the case of a fluid but these are all derived from the scalar, which determines the nature of the wave).
2. Pressure waves in a fluid are often referred to as *longitudinal* ones. This is a practice adopted by convention, since plane waves in a fluid (as also spherical waves) can be described in terms of the longitudinal strain (i.e., strain along the direction of wave normal, which is the direction of propagation for plane and spherical waves) as the wave function (see sec. 9.11.2). A more logical approach of classifying waves is to describe them as ones corresponding to *scalar*, *vector*, or *tensor* wave functions.
3. A few words are in order for sound waves in an *anisotropic* solid where it is assumed that it can be looked upon as a continuous medium. Considering once again a plane wave with a wave vector \mathbf{k} and frequency ω , it is found that there can be *three* independent waves with these values of \mathbf{k} and ω . In general, none of these waves can be described as purely longitudinal or transverse. This, of course, relates to the tensorial nature of these waves.
4. When one takes into account the discreteness of the atomic units constituting a crystalline solid, one finds that, in general, the number of independent waves for a given \mathbf{k} and ω is more numerous than that in the case of a continuous medium. In contrast to three independent sound waves in a continuous medium, the additional types of waves that can propagate in a discrete medium are termed *optical* waves - a nomenclature that does not imply that these are of the same

nature as light waves.

In this context, mention may be made of *longitudinal* waves set up in a thin rod made of an isotropic solid. These waves can be set up by drawing along the length of the rod near one of its ends with a piece of leather, where the rod is held fixed by means of one or more clamps. It is such an arrangement that has been shown in fig. 9-34 for setting up standing waves in air in a closed tube. The velocity of longitudinal waves in such a thin rod, whose thickness is small compared to the wavelength of the longitudinal wave, is given by the expression

$$v = \sqrt{\frac{Y}{\rho}}, \quad (9-80)$$

where Y stands for the Young's modulus of the solid and ρ for its density.

Problem 9-11

The tension in a stretched metal wire is such as to cause a tensile strain $\epsilon = 10^{-4}$ in it. If the Young's modulus of the material of the wire is $Y = 2.0 \times 10^{11} \text{ N}\cdot\text{m}^{-2}$ and its density is $\rho = 8.0 \times 10^3 \text{ kg}\cdot\text{m}^{-3}$, find the velocity of longitudinal elastic waves through the wire, and the velocity of transverse vibrations propagating through it.

Answer to Problem 9-11

HINT: According to formula (9-80), the velocity of a longitudinal elastic wave through the wire is $v_{\text{long}} = 5.0 \times 10^3 \text{ m}\cdot\text{s}^{-1}$. Further, looking at the formula for the velocity (v_{trans}) of transverse vibrations propagating along the stretched string as given by eq. (9-73), and comparing it with the formula (9-80), one finds that $v_{\text{trans}} = \sqrt{\epsilon} v_{\text{long}}$, where ϵ stands for the tensile strain produced in the wire by the tension stretching it (check this out by making use of the stress-strain relation in the wire). One thereby gets $v_{\text{trans}} = 50.0 \text{ m}\cdot\text{s}^{-1}$.

Finally, I may also mention *bending* and *torsional* waves in a thin rod where both are in the nature of transverse waves. The velocities for these waves differ from the shear

waves in an infinitely extended medium. Longitudinal waves and transverse bending waves can also be set up in thin plates. I do not include here the expressions for the velocities of these various waves in rods and plates.

Chapter 10

Ray Optics

10.1 Introduction

Optics is the subject dealing with phenomena relating to the production, propagation, and recording of light. The description and explanation of a large class of such phenomena is based on *electromagnetic theory*, also known as the *classical* electromagnetic theory, which I will briefly introduce in chapter 14. According to this theory, light emitted from a source propagates in the form of electromagnetic waves, i.e., waves made up of oscillations of electric and magnetic field intensities at various points of space. Such a wave, in the course of its propagation may encounter various objects like, say lenses, mirrors, apertures, etc., that modify the wave which is then recorded by an appropriate device such as the human eye. This approach of looking at optical phenomena is commonly referred to as *wave optics*.

Two other approaches in optics are best introduced in their relation to wave optics. One of these, the *quantum theory* makes use of the *photon* picture of the electromagnetic field, describing the latter from the quantum point of view. It constitutes, in the present state of knowledge, the ultimate theory of light with reference to which the classical electromagnetic theory can be considered to be an approximate scheme of description and explanation which, however, is a vastly useful one. While the classical theory explains a large class of optical phenomena, it does have its limitations in certain contexts involv-

ing the interaction of light and matter, where the more fundamental quantum theory proves to be of value.

The other theory dealing with optical phenomena, which I will introduce in the present chapter, is *ray optics*, where the term *geometrical optics* is also in common use because, in this branch of optics, the propagation of electromagnetic waves is described in terms of a number of geometrical principles. In a sense, ray optics bears a similar relation to wave optics as wave optics itself does to quantum optics. More precisely, ray optics can be looked upon as an approximation scheme, deriving from wave optics, for describing the propagation of light, and is applicable in a limited set of contexts while being at the same time a simple and familiar approach in optics.

Later in this chapter, as also in chapter 15 I will dwell more on the way ray optics relates to wave optics, where a brief outline of electromagnetic theory (chapter 14) will prove to be useful. The present chapter outlines the basics of ray optics without direct reference to wave optics.

10.2 Ray optics: basic principles

Ray optics is built up on a number of basic working principles, most of which can be traced back to classical electromagnetic theory. In the present section, I will list a number of these working principles, which will then be made use of in subsequent sections so as to put together an outline of ray optics.

1. A source of light emits energy in the form of what we will refer to as an 'optical disturbance' where, in reality, an optical disturbance is nothing but an electromagnetic wave propagating from the source to distant regions of space. More precisely, the electric and magnetic field intensities at any given point in space vary with time, and these variations are transmitted from one region of space to another.

The energy carried by the optical disturbance proceeds from one point of space to another along a path referred to as a *ray*. In most of the present chapter, the optical disturbance will be assumed to be emitted by a *point source*, while an

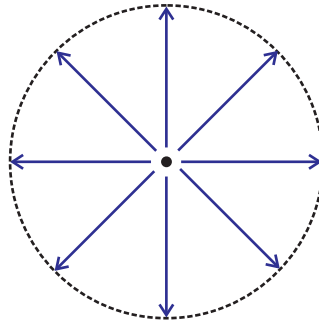


Figure 10-1: Illustrating the emission of energy along rectilinear ray paths from a point source; the source generates an optical disturbance that carries the energy to distant regions away from it.

extended source will be assumed to be made up of a large number of independent point sources. A point source is a convenient and useful idealization in optics.

Energy is emitted from a point source in the form of rays directed *radially* away from the source, as in fig 10-1.

2. The propagation of energy carried by an optical disturbance in a *homogeneous medium* takes place along *rectilinear* ray paths, as in fig. 10-2.
3. On analyzing an optical disturbance, it is, in general, seen to be made up of a number of disturbances of a relatively simple nature, namely, *monochromatic waves*. A monochromatic wave is characterized by some specific value of *frequency* or, alternatively, of *wavelength*.

The speed of propagation in a given medium of an optical disturbance carried by a monochromatic wave of a specified frequency is determined by a certain characteristic of the medium referred to as its *refractive index*. For any given medium, the refractive index depends on the frequency of the monochromatic wave under consideration (see sec. 14.7.1) and, in general, increases with the frequency. The variation of refractive index with frequency is referred to as *dispersion*.

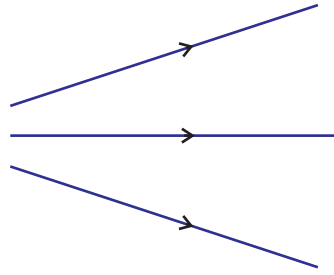


Figure 10-2: Illustrating the propagation of energy, carried by an optical disturbance, along rectilinear ray paths in a homogeneous medium; the refractive index is the same at all points in the medium; the propagation of energy along a ray takes place in the direction of the arrow.

Considering any particular monochromatic wave, if n be the refractive index (also termed the *absolute* refractive index) of the medium corresponding to the frequency of the wave, then its speed v (the phase velocity, see sec. 14.4.2) in the medium is given by

$$v = \frac{c}{n}, \quad (10-1a)$$

where c stands for the velocity of light *in vacuum*. Indeed, an optical disturbance can propagate in vacuum, i.e., even in the absence of any material medium, and the speed of propagation in vacuum is *the same for all frequencies*, being $c = 3 \times 10^8$ m·s⁻¹ (approx).

An alternative way to write the above equation is

$$n = \frac{c}{v}. \quad (10-1b)$$

The frequency (ν) of a monochromatic wave relates to the rate at which the electric and magnetic field intensities at any given point of space vary with time. The concept of wavelength (commonly denoted by λ) , on the other hand, relates to the manner in which the electric and magnetic field intensities vary in space at any given instant of time. The two characteristics of a monochromatic wave in a

medium are related to each other as

$$\nu\lambda = v, \quad (10-2)$$

which implies that the wavelength varies inversely as the frequency: waves of relatively larger frequencies have smaller wavelengths and, conversely, smaller frequencies imply larger wavelengths. The frequency of an electromagnetic wave can be anything from zero to infinity. Not all of these possible values, however, correspond to *visible light*. The latter corresponds to a range of frequencies from around $3.7 \times 10^{14} \text{ s}^{-1}$ to around $1.0 \times 10^{15} \text{ s}^{-1}$. Waves of different frequencies in this range generate the sensation of various *colors* in the eye. While the frequencies of electromagnetic waves from very small to very high values are said to make up the *electromagnetic spectrum*, the frequencies corresponding to visible light mentioned above make up the *optical spectrum*. The lower frequency end of the optical spectrum generates the sensation of red color, while the upper frequency end generates the sensation of violet.

A number of concepts relating to wave motion have been introduced in chapter 9. Electromagnetic and optical waves will be introduced in greater details in chapters 14, 15.

The velocity with which the surfaces of constant phase of a monochromatic wave (refer to sections 9.4.2, 14.4.4) propagate through a medium is, in general, different from the velocity with which energy is transported through the medium along the ray paths. While the former is termed the *phase velocity* of the wave, the latter is referred to as the *group velocity*. The group velocity also gives, under ordinary circumstances, the velocity of propagation of an optical *signal* through the medium under consideration. Optical signal propagation is relevant in modern day optical *communications* systems.

4. In practice, light rays, in course of propagation, encounter *inhomogeneities* in the form of a variation of the refractive index, in which case the ray paths deviate from straight lines. Two instances of ray paths in the presence of inhomogeneities are shown in fig. 10-3(A) and (B). In (A), a ray is seen to change its path while entering

from one medium to another, the refractive indices of the two media being different. This phenomenon of bending of ray path as an optical disturbance enters from one medium to another is referred to as *refraction*. At the interface separating the two media, part of the optical disturbance gets sent back to the medium where it came from, following a different path. This is the phenomenon of *reflection*.

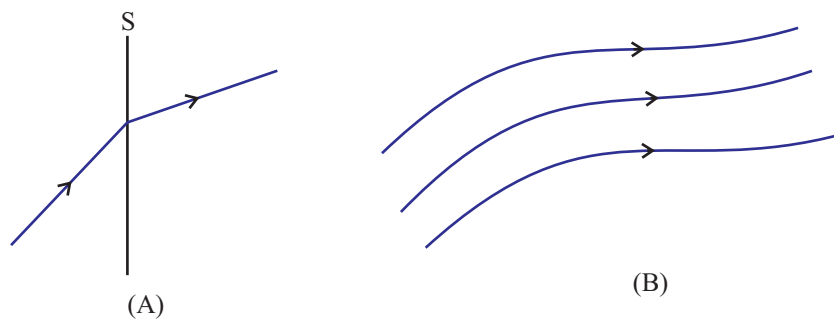


Figure 10-3: Bending of a ray path caused by inhomogeneities; (A) as a ray encounters a surface S (section by the plane of the figure shown) separating two media with different refractive indices, it changes its course and gets bent by refraction (the reflected ray is not shown); (B) a bunch of curved ray paths in a medium with a continuously varying refractive index.

Reflection and refraction of light can be looked at as special instances of corresponding phenomena observed in the context of wave motion in general and of the propagation of electromagnetic waves in particular. For background material relating to reflection and refraction of waves, see sections 9.8 and 14.6; see also sec. 15.9 where the relation between wave optics and ray optics is reviewed.

While in fig. 10-3(A), the inhomogeneity arises in the form of an abrupt change in the refractive index at the interface separating two homogeneous media, fig. 10-3(B) shows a bunch of ray paths in a medium where the refractive index varies *continuously*, or gradually, from one point to another in a medium. Here the rays are seen to follow curved paths along which the energy carried by the optical disturbance propagates.

The central task in ray optics is to determine the ray paths as an optical dis-

turbance, originating from a source, travels from one region of space to another (through an optical fiber, for instance; in this book I will not, however, tell you anything more of the immensely important subject of *fiber optics*), encountering various inhomogeneities.

A principle of fundamental importance in this context, known as *Fermat's principle*, is a general one that can be made use of so as to determine the ray paths in various different situations including those involving media with continuously varying refractive indices. In this book, however, we will not consider such general problems in ray optics and instead will concentrate on situations where a ray path is made up of segments, each segment being a straight line in a homogeneous medium (fig. 10-4).

As a ray path, following a straight line segment, encounters an interface separating two different media, it undergoes reflection or refraction, giving rise to a new segment along a different direction. In general, a ray undergoes *both* reflection and refraction at an interface, giving rise to *two* new segments. Depending on the problem under consideration, one often has to consider only one of the two segments for the determination of the required course followed by the ray path.

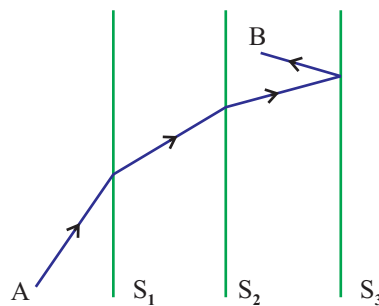


Figure 10-4: Illustrating a ray path made up of straight line segments; the ray encounters surfaces S_1 , S_2 , S_3 (sections by the plane of the figure shown), where each of these surfaces separates two media with different refractive indices; while at each of these surfaces there arises a reflected and a refracted ray, the figure shows only the refracted rays at S_1 and S_2 , and the reflected ray at S_3 which are relevant for the ray course from point A to point B.

To determine how a segment of ray path gets bent in reflection and refraction, we will make use of the *laws of reflection and refraction* - two familiar and widely used principles. These are simpler to state compared to Fermat's principle, being, in a sense, special instances of the latter.

5. *The laws of reflection.* Fig. 10-5 shows a surface S separating two media with different refractive indices on which a ray AP is incident at the point P. The ray is reflected back into the same medium along the path PB. PN is the normal to the surface S at the point of incidence P. The path followed by the reflected ray PB for a given incident ray AP can be determined by making use of the following two principles referred to as the laws of reflection.

(a) The incident ray, the reflected ray, and the normal to the reflecting surface (S in fig. 10-5), all lie in one single plane (shown with dotted lines in fig. 10-5).

(b) The angle of incidence (i.e., the angle between the incident ray and the normal) and the angle of reflection (i.e., the angle between the reflected ray and the normal) are equal to each other. In other words, referring to fig. 10-5,

$$i = i'. \quad (10-3)$$

I will later indicate how this formula is to be modified so as to take into account the *sign convention in ray optics*.

6. *The laws of refraction.* Fig. 10-6 shows a surface S separating two media with different refractive indices on which a ray AP is incident at the point P from the medium to the left of S. The ray is refracted into the medium to the right of S along the path PB. NPN' is the normal to the surface S at the point of incidence P. The path followed by the refracted ray PB for a given incident ray AP can be determined by making use of the following two principles referred to as the laws of refraction.

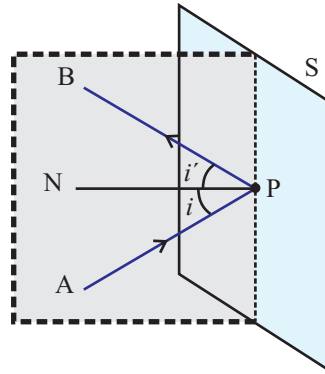


Figure 10-5: Illustrating the laws of reflection; the incident ray path AP, the reflected ray path PB, and the normal PN to the reflecting surface S, all lie in one plane; the angle of incidence i and the angle of reflection i' are equal.

(a) The incident ray, the refracted ray, and the normal to the surface of separation (S in fig. 10-6), all lie in one single plane (shown with dotted lines in fig. 10-6).

(b) The angle of incidence (i.e., the angle between the incident ray and the normal) and the angle of refraction (i.e., the angle between the refracted ray and the normal) are related to each other as (see fig. 10-6)

$$n_1 \sin i = n_2 \sin r, \quad (10-4a)$$

where n_1 and n_2 are the refractive indices of the two media respectively. The above formula is commonly referred to as *Snell's law*.

An alternative way to write the above formula is

$$\frac{\sin i}{\sin r} = \frac{n_2}{n_1} = n \text{ (say)}, \quad (10-4b)$$

where $n(= \frac{n_2}{n_1})$ is termed the *relative* refractive index of the second medium with respect to the first.

(a) The laws of reflection and refraction hold even when the surface separating the

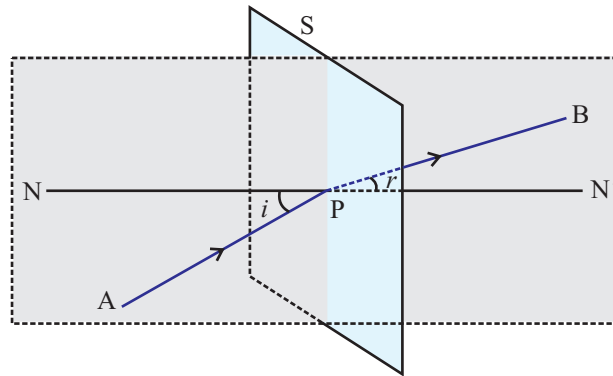


Figure 10-6: Illustrating the laws of refraction; the surface S separates two media with refractive indices n_1 and n_2 ; the incident ray path AB in the first medium gets bent along PB into the second medium, while NPN' is the normal to the surface S at the point P ; all the three lines lie in one single plane and, moreover, the angle of incidence i and the angle of refraction r satisfy (10-4a).

two media is a curved one, the normal to the surface at the point of incidence being, in that case, perpendicular to the tangent plane to the surface at the point of incidence. However, the conditions of validity of these laws are that the radius of curvature of the surface is to be large compared to the wavelength of the incident light and, at the same time, the extent of the surface is to be large compared to the wavelength (see sec. 9.8 for necessary background). Assuming that these conditions are satisfied, we will make use of these laws to study *image formation* due to reflection and refraction at spherical surfaces.

- (b) Looked at from the point of view of wave optics, the formation of images can be interpreted as, roughly, a concentration of electromagnetic energy at some point or set of points in space, where this concentration of energy may assume the form a *singularity* of the electromagnetic field variables. More generally, the concentration of energy may not be ideal because of *aberrations* and of *diffraction* effects. In this chapter, however, we will consider image formation on the basis of a set of idealized rules, namely the rules of geometrical optics, where the image of an object can be described in terms of a number of geometrical relations. In this sense, the images considered in ray optics are referred to as *geometrical images*.

7. *The law of intensity ('intensity rule' of geometrical optics).* The *intensity* at any point P due to an optical disturbance is defined as follows:

Considering the ray path passing through P, imagine a small surface of area δA around P, with its normal directed along the ray path (see fig. 10-7). Let the energy, carried by the optical disturbance, crossing this area per unit time be δW . Then the ratio $\frac{\delta W}{\delta A}$, i.e., the energy crossing per unit area per unit time is termed the intensity at P. In this definition, an *averaging* with respect to time is implied, over an interval large compared to the time period of the optical disturbance under consideration (refer to sec. 9.11.4).

The intensity as defined above has an important visual effect. If a screen be held in the path of the optical disturbance, then those points on the screen where the intensity is high will appear bright while other points at which the intensity is relatively low will appear dark by comparison.

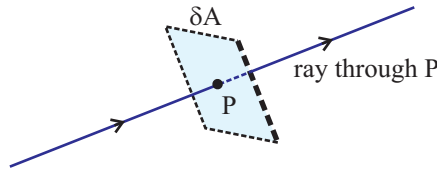


Figure 10-7: Illustrating the definition of intensity at a point; the intensity at P is the time-averaged rate of flow of energy per unit area along the direction of the ray passing through P.

Looking at fig. 10-1, where a bunch of rays diverging from a point source is shown, the *intensity rule* states that the intensity at any point at a distance r from the source is proportional to $\frac{1}{r^2}$.

According to this rule, the intensity attains a high value as one approaches the point source, tending to infinity as r approaches zero (we assume that the source emits energy at a finite rate). In reality, however, the intensity cannot be infinitely large since the source of light cannot be an ideal point source while, at the same time, emitting energy at a finite rate.

Consider now a situation that is, in a sense, opposite to that in fig. 10-1, namely, one where a bunch of rays *converges* at a point P, as in fig. 10-8. Here again, if one

approaches the point P along any one of these rays, the intensity will attain large values, being infinitely large at the point P itself.

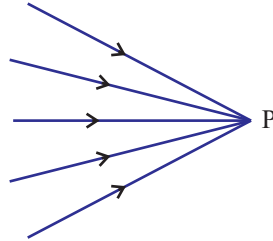


Figure 10-8: A bunch of rays converging toward a point P; the intensity becomes large as one approaches P along any of the ray paths, being infinitely large at the point P itself.

In reality, however, the intensity does not attain an infinitely large value because of the wave nature of light which results in a spreading out of the energy carried by the optical disturbance.

The fact that the intensity at a point like P in fig. 10-8 can be very large, is relevant in the context of *image formation* in ray optics. This I turn to in the next section.

8. *Polarization rules.* The basic principles of geometrical optics also imply a set of *polarization rules* that tell us how the state of polarization of a signal made of linearly polarized light (see section 14.4.8 for background) changes along the path of a ray. In the present chapter, however, I will make no reference to the states of polarization of the light rays.

Together, these few principles make up the entire basis of the subject of ray optics.

Digression: negative refractive index.

Though the refractive index is a frequently used parameter in ray optics, the physical basis of the refractive index of a material relates to the interaction of *electromagnetic waves* with the constituents of the medium. All the commonly known materials are characterized by positive values of the refractive index. Artificially fabricated materials have, however, been developed in

recent decades for which the refractive index is effectively *negative*. These belong to a broad class of artificial materials that have quite exotic properties and are known as *metamaterials*. The negative refractive index of certain metamaterials is of great importance from the point of view of potential applications. See sec. 14.7.3 for a brief introduction to the negative refractive index metamaterials.

10.3 Image formation by rays originating from a point source

Fig. 10-9 shows a bunch of rays diverging from a point source O in a medium A. Assuming that A is a homogeneous medium, the rays fan out in all directions along straight line paths, a number of such rays being shown in the figure. Such a bunch of rays is termed a *divergent* beam. Suppose that the rays are incident on a smooth surface S separating the medium A from another medium A'. In this book, we will consider only plane and spherical surfaces at which the rays undergo reflection and refraction. The rays originating from O are seen to be refracted at the surface S into the medium A' where they form a different bunch. The bunch of rays in the medium A' shown in the figure is an instance of a *convergent* beam. However, the figure is to be taken as only an illustration - it is not always that a divergent beam is converted into a convergent one by refraction at a surface.

The bunch of rays under consideration may undergo further changes of course by refraction and reflection at other surfaces, but these are not shown in fig. 10-9. A set of successive surfaces like S forms what is referred to as an *optical system*. Suppose that the bunch of rays under consideration passes through such an optical system and comes out in the form of a convergent beam into the medium B on being refracted at the surface S' as in fig. 10-9. This beam finally converges to the point I in the medium B, whereafter it diverges again.

If viewed from the other side of I, the rays diverging from I appear as if these are emitted from a point source, similar to the rays coming out from O. This fact is expressed by

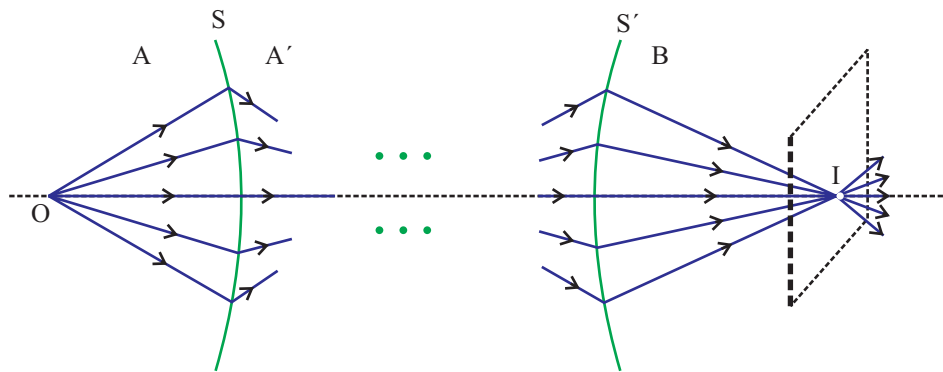


Figure 10-9: Image formation by rays originating from a point source; a divergent beam incident on the surface S separating two media A and A' is refracted into A' as a convergent beam; the beam changes its course by reflection or refraction at other surfaces (not shown, represented by dots); after passing through the optical system made up of all the surfaces, the beam emerges into the medium B as a convergent one, finally converging to the point I , which is the real image of O formed by the optical system.

saying that I is the *image* of O formed by the optical system under consideration. If, instead of being viewed from the other side, a screen is placed at I (dotted surface in fig. 10-9) then a bright spot will be formed on the screen at I , other points around I being dark by comparison. This is expressed by saying that I is the *real* image of the point object O .

Another instance of image formation is shown schematically in fig. 10-10. Here again the bunch of rays diverging from the point source O pass through an optical system, emerging finally in the medium B , but now as a *divergent* beam. These rays do not meet at a point as in fig. 10-9. However, on being produced backwards (dotted lines in the figure) they meet at the point I . In this case, if one imagines the medium B to be extended to the other side of S' and a point source of light to be placed at I , then the rays that would have been emitted from I , when viewed from the right of S' would appear similar to those actually emerging into the medium B as in the figure.

In other words, with reference to the divergent beam to the right of S' , I acts like a point source. Thus, I is here the image of the point source of O , but of a different kind compared to that in fig. 10-9. It is termed a *virtual* image, which cannot be captured on a screen since the rays do not actually converge at I .

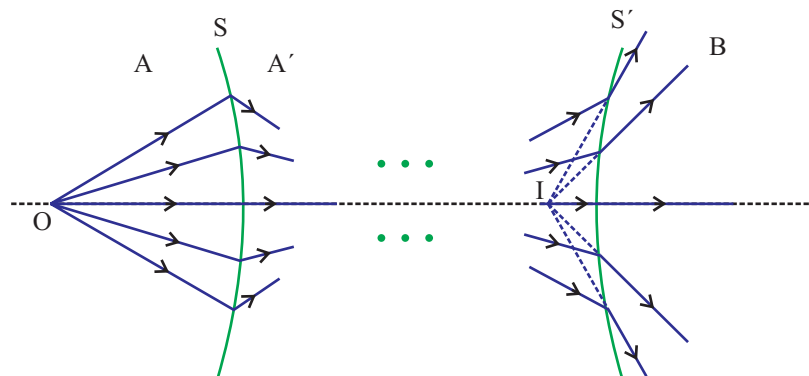


Figure 10-10: Image formation by an optical system, as in fig. 10-9, where now the image I is a virtual one; when viewed from the right of S' , the rays appear to have originated from a point source; the image, however, cannot be captured on a screen since the rays emerging into the medium B do not actually meet at I .

While we have considered in figures 10-9 and 10-10 the formation of point images for point objects, in reality, optical systems are used to form images of *extended* objects. Such an extended object may be looked upon as a collection of so many point objects, each giving rise to a point image formed by the optical system under consideration. All these point images, real or virtual, make up an extended image of the object.

There is, however, an important qualification that I must add relating to the formation of a point image of a point object. In reality, a point object does *not*, in general, give rise to a point image. What actually happens can be explained with the help of fig. 10-11, which gives a schematic view of the final emergent beam tending to converge to an image point. As can be seen in the figure, the beam does not, however, converge to a single point. Instead, the rays come close to one another and then fan out in the form of a divergent beam. The rays marked '1', '2', and '3' are seen to meet at the point I_1 while the rays marked '4' and '5' meet at a different point I_2 . Evidently, here a concentration of energy at a single point does not take place (which, in any case, is ruled out by the wave nature of light), and a point image is not formed. Instead a small region of non-zero extent but of considerable brightness will be formed near the points I_1 and I_2 . This is referred to as an image with an *aberration*, where the term aberration means an error, or an anomaly, in image formation. The smaller the extent of the bright region, the more nearly perfect will the image be. While the figure shows how an aberration can

arise in the formation of a real image, aberrations can similarly arise in the formation of a virtual image as well.

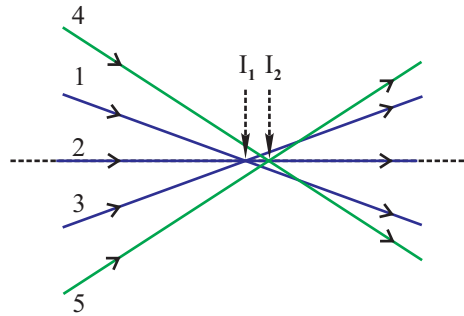


Figure 10-11: Aberration in the formation of a real point image; the rays emerging from the optical system (not shown) fail to meet at a single point; the rays marked '1', '2', and '3' meet at I_1 while those marked '4' and '5' meet at I_2 ; instead of a point image one has here a small but extended bright region.

In reality, the image formation by any optical system involves aberrations to some extent. However, one can lessen the aberrations by an appropriate designing of the system.

One can, thus, speak of two sources of imperfection whereby a point object fails to produce a point image: first, the spreading out of the optical disturbance due to the wave nature of light and, next, the aberrations in optical systems due to which a convergent bunch of rays does not actually converge at a point or a divergent bunch does not appear to fan out from a single point when produced backward. Of these, the aberration effects can be minimized by adopting special measures, while the spreading out of the wave disturbance is of more fundamental nature.

I now turn to a consideration of instances of reflection and refraction at plane and spherical surfaces and of image formation by one or more such surfaces.

10.4 Image formation by reflection at a plane surface

You may already be familiar with features of reflection at a plane surface. Figure 10-12 depicts image formation of a point object by reflection at a plane surface. A divergent

bunch of rays originating from the point source O changes its course on reflection from a plane surface S (see fig. 10-12). The reflected ray paths, on being produced backwards, meet at the point I which is the virtual image of O . On looking at the reflected rays, they appear to have originated from I though, in reality, no energy actually gets concentrated at or is emitted from I . The image I cannot be captured on a screen or a photographic film.

Here I do not refer to the refraction occurring at the surface S , along with reflection. On applying a deposit of an appropriate reflecting layer on the surface its *reflection coefficient* can be made to attain a value close to unity, most of the energy incident on the surface being then reflected back from it, with only little energy transmitted into the other side of the surface. The surface then acts as a *reflector* or a *mirror*.

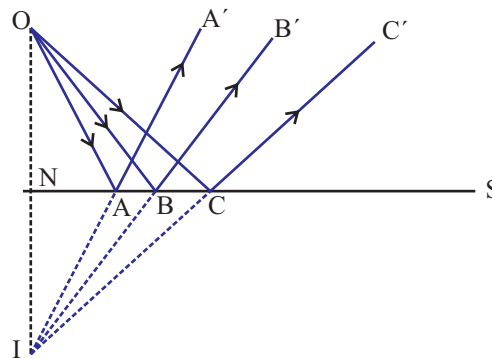


Figure 10-12: Image formation by reflection at a plane surface S ; rays diverging from the point source O are reflected at the surface, and the ray paths on being produced backwards intersect at I , which is thus a virtual image of O .

One other important feature of the image formed by reflection at a plane surface is that *it is free of aberrations*. Indeed, the image point I can be characterized as follows. Let a perpendicular be dropped from O onto the surface S , the foot of the perpendicular being N . Let the segment ON be then produced up to I such that $ON=NI$. The point I so obtained then gives the position of the image, and *all* the reflected rays (AA' , BB' , CC' ,...), on being produced backwards meet at I , regardless of where the points of incidence are.

Problem 10-1

Establish the validity of the above statement.

Answer to Problem 10-1

Consider an arbitrarily chosen incident ray OA, and the corresponding reflected ray AA' in fig. 10-12. Show that A'A, on being produced backwards, passes through I as obtained in the above construction. Equivalently, show that, if IA is produced then the segment AA' so obtained is the reflected ray path, obeying the laws of reflection. For this, note that the normal ON, the incident ray path OA, and the segment AA' lie in the same plane. Moreover, the triangles OAN and IAN being equal in all respects, $\angle OAN = \angle IAN$.

This is one of the rare instances in optics where a point object gives rise to a point image *without* aberrations. If now an extended object is considered and looked upon as a collection of point objects, then the corresponding point images make up an extended image which is an exact replica of the object in the sense that, if it is reflected a second time, then the second image can be made to coincide exactly with the object by a parallel translation.

Problem 10-2

Images by multiple reflection

Fig. 10-13 shows two plane mirrors OM₁, OM₂, inclined at an angle θ to each other, the planes of the mirrors being perpendicular to that of the figure. P represents a point object, of which images are formed by multiple reflections in the two mirrors. Angles are measured from OM₁ chosen as the reference line, where $\angle POM_1 = \phi$ (thus, $\angle M_2OP = \theta - \phi$). Rays from P, suffering their first reflection at OM₁, form the image A₁, where $\angle M_1OA_1 = -\phi$, and $OP = OA_1$ (check this out; thus P and A₁, as also the other images all lie on a circle C with O as center).

A₁ acts as an object for the mirror OM₂, i.e., the rays originating from P and first reflected from OM₁ suffer the next reflection from OM₂, and form the image A₂, where A₂ lies on the circle C. In turn, A₂ acts as an object for OM₁ and forms the image A₃. This continues, and one gets a series

of images A_k ($k = 1, 2, \dots$), all resulting from rays reflected first from OM_1 , and all lying on the circle C . The series ends with an image formed within the arc $M'_1M'_2$ lying behind both the mirrors, in which case no further reflection is possible (reason this out). In a similar manner, a second series of images B_k ($k = 1, 2, \dots$) is formed on C , by rays first reflected from the mirror OM_2 .

Let, for any given value of k , the angle made by OA_k with the reference line OM_1 be ϕ_k , and that made by OB_k be ψ_k . Obtain formulas for ϕ_k , ψ_k .

Answer to Problem 10-2

HINT: Evidently, $\phi_1 = -\phi$, $\phi_2 = 2\theta + \phi$. These formulae tell us how an initial angle gets transformed by reflection at the two mirrors. Thus, substituting $-\phi$ for ϕ , one gets, from the second of the above two formulae, $\psi_1 = 2\theta - \phi$ (check this out), and then substituting $2\theta - \phi$ for ϕ in the first formula of the two, $\psi_2 = -2\theta + \phi$. Applying these transformations repeatedly, one gets

$$(k \text{ odd}) \phi_k = -(k-1)\theta - \phi, \psi_k = (k+1)\theta - \phi; (k \text{ even}) \phi_k = k\theta + \phi, \psi_k = -k\theta + \phi.$$

.

Digression: total number of images

One can go on from where we arrived at in problem 10-2 and work out the total number of images of P formed by the two mirrors inclined to each other by the angle θ . This, however, requires a bit of involved reasoning, and the result does not admit of a single neat expression since there occur a number of degenerate, or exceptional, situations for which the number differs from that given by the general expression, where the latter holds for a non-degenerate situation. To be more precise, if θ is not a sub-multiple of 2π (in which case it is also not a sub-multiple of π), and at the same time θ is not a sub-multiple of $\pi - \phi$ or $\pi + \phi$ (these are the cases respectively when the last image of the series $\{A_k\}$ or $\{B_k\}$ coincides with M'_1 or M'_2), the number of images is given by the expression

$$N = \left[\frac{\pi - \phi}{\theta} \right] + \left[\frac{\pi + \phi}{\theta} \right] + 1.$$

In this expression, the symbol $[\cdot]$ stands for the *integer part* of the argument.

For instance, with $\theta = \frac{26\pi}{180}$ and $\phi = \frac{5\pi}{180}$, one has $N = 14$ while, with $\theta = \frac{26\pi}{180}$ and $\phi = \frac{\pi}{180}$, $N = 13$.

Interesting degenerate situations correspond to θ being a sub-multiple of π ($N = \frac{2\pi}{\theta} - 1$), and θ

being a sub-multiple of 2π but not of π ($N = \frac{2\pi}{\theta}$; however, the case $\phi = \frac{\theta}{2}$ is excluded here). Thus, for $\theta = \frac{\pi}{3}$, $N = 5$ and, at the same time, for $\theta = \frac{2\pi}{5}$ ($\phi \neq \frac{\pi}{5}$), $N = 5$.

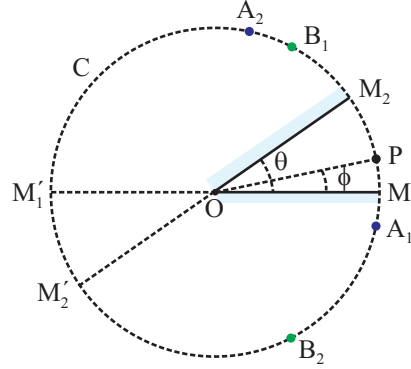


Figure 10-13: Series of images formed of a point object P by a pair of plane mirrors inclined to each other at an angle θ ; the images are formed on a circle C with O as center; the images can be grouped into two series - the series of images A_1, A_2, \dots , is formed by rays first reflected in the mirror OM_1 , while the series B_1, B_2, \dots , is formed by rays first reflected from OM_2 ; the images in either series are formed by successive reflections in the two mirrors till an image is formed on the arc $M'_1M'_2$ behind both the mirrors; the total number of images formed depends, in general, on θ as also on the angle ϕ ; however, there exist interesting degenerate cases when the number does not depend on ϕ .

10.5 Refraction at a plane surface

10.5.1 Image formation

In fig. 10-14, S is a plane surface (section by the plane of the figure) separating two media of refractive indices n_1 and n_2 , where O is a point source located in the first medium. The ray ON originating from the source is incident normally on the surface of separation at N and, in accordance with Snell's law, enters into the second medium along NA , the latter being an extension of ON (check this statement out). A second ray OQ is also shown in the figure, incident on the surface of separation S at Q , which we assume to be a point *close to* N . On refraction at the surface, the ray enters into the second medium along the path QB which, on being produced backwards, intersects ON at I . On making use of Snell's law, and the fact that the distance NQ ($= d$, say) is small,

one obtains

$$\frac{v}{u} = \frac{n_2}{n_1}, \quad (10-5)$$

where u stands for the distance (ON) of the object point from the surface S and v similarly denotes the distance of the point I (obtained as above; being also the image point of O, as we see below) from S.

Problem 10-3

Check the validity of the above statement relating to eq. (10-5).

Answer to Problem 10-3

HINT: Let i and r be the angles of incidence and refraction for the ray OQ, as shown in the figure, MQM' having been drawn normal to the surface S at Q. One then finds,

$$\frac{v}{u} = \frac{\cot r}{\cot i} = \frac{n_2 \cos r}{n_1 \cos i}, \quad (10-6)$$

where eq. (10-4a) has been made use of.

In this equation, the angles i and r are small since the distance $d(= NQ)$ is assumed to be small (more precisely, we assume that $d \ll u$ in the present instance) and hence one has $\cos i \approx 1$, $\cos r \approx 1$. This gives eq. 10-5. In this equation, second degree terms in $\frac{d}{u}$, and terms of higher degrees have been ignored.

Note that eq. (10-5) does not involve the distance d , which implies that the point of intersection N is the *same for all rays* (such as, say, OQ and OQ' shown in fig. 10-14). This means that I is the virtual image point of O formed by refraction at the plane surface S.

However, unlike the case of image formation by reflection, this image is formed *only* by the rays whose points of incidence are sufficiently close to N. Such rays are referred to as *paraxial rays* where, in the present context, the ray path ONA can be looked upon as

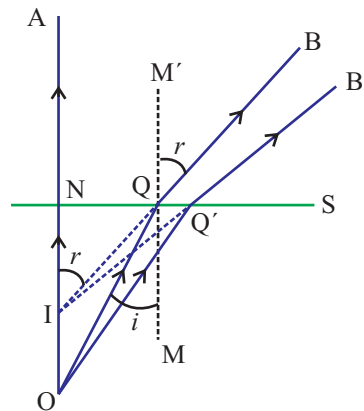


Figure 10-14: Image formation (schematic) by refraction at a plane surface S (section by the plane of the figure) separating two media; rays diverging from the point source O are refracted at the surface, and the ray paths on being produced backwards intersect at I , which is thus a virtual image of O ; of the three ray paths shown, one (ONA) is normal to the surface; the points of incidence of the rays on S are assumed to be *close to one another*, and hence close to N ; when a refracted ray path is produced backwards its point of intersection with ON is seen to be independent of the point of incidence on S and is thus the same for all rays sufficiently close to ONA ; OQ and OQ' are referred to as *paraxial rays*.

the axis. Rays originating from O but incident on S at points not close to N are refracted along paths which do not connect to I on being produced backwards. In other words, when looked at from above the surface S , *all* ray paths do *not* appear to originate from I . This results in *aberrations* affecting the image whereby the latter fails to be a perfect replica of the object.

Figure 10-15 shows a number of *non-paraxial* rays, ones whose points of incidence are not sufficiently close to N . Among these rays, OR_1C_1 , OR_2C_2 , and OR_3C_3 are ray paths for which the points of incidence (R_1, R_2, R_3) are close to one another, but distant from N . The refracted rays belonging to this bunch, when produced backwards, meet at I' , which is seen to be distinct from I , the image formed by the paraxial rays (refer to fig. 10-14). I' is thus the image formed by the bunch of non-paraxial rays under consideration. Similarly, considering another set of non-paraxial ray paths OS_1D_1 , OS_2D_2 , and OS_3D_3 whose points of incidence (S_1, S_2, S_3) are close to one another, the image I'' formed by these rays differs from both I and I' . Thus, considering all these bunches, there does not result a single point image for a point object, in contrast to the case of reflection at

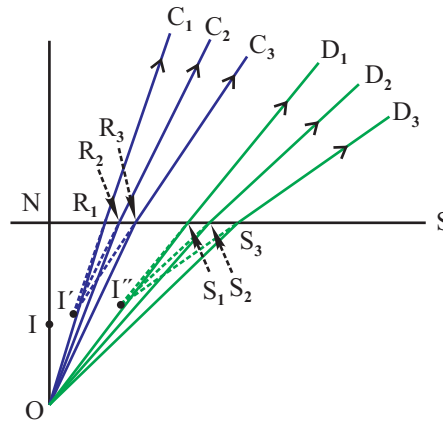


Figure 10-15: A set of non-paraxial rays suffering refraction at a plane surface S (compare with fig. 10-14); two such bunches are shown; the refracted ray paths for one of the bunches, on being produced backwards, intersect at I' , while the corresponding point for a second bunch is I'' , different from I' ; each bunch is formed by rays close to one another; thus I' , I'' , etc., are images formed by the various different bunches, these images being all distinct from the image I formed by the paraxial rays (not shown); thus, instead of a point image for the point object O , there is formed an illuminated surface (not shown) made up of all these various 'image' points; this is an instance of aberration in image formation.

a plane surface. Instead, an *extended* image is formed, made up of the images I , I' , I'' , etc, of which I is referred to as the *paraxial image*. This is an instance of aberration in image formation. If one confines oneself to a consideration of the paraxial rays alone, then one obtains a point image (I) for a point object (O).

The type of aberration encountered here is termed *monochromatic* aberration because it arises even when the optical disturbance emitted by the object O is a monochromatic one, i.e., is carried by a wave characterized by a definite frequency and a definite wavelength. In contrast, another type of aberration arising in the image formation by an optical system is referred to as *chromatic* aberration, which is a defect in image formation resulting from the optical disturbance emitted from the source being a mixture of monochromatic waves with a number of distinct frequencies.

In eq. (10-5), u and v are termed the *object distance* and *image distance* respectively. It is important to take note of the fact that these are, in reality, *distances with appropriate signs*, i.e., their numerical values carry a magnitude and a sign (+ or -). Indeed, the mathematical relations between various distances and angles that one writes in ray

optics, all require appropriate signs to be affixed to these quantities, in the absence of which the relations lead to inconsistent results.

The quantities that are to be assigned appropriate signs in ray optics can be grouped into three classes, namely, axial distances, transverse distances, and angles. Choosing an appropriate straight line as axis for the optical system under consideration (for instance, the line ON in fig. 10-14 or 10-15), and any one of the two possible directions along that straight line as the positive direction (say, the upward direction, i.e., from O towards N in each of these two figures) one has to affix positive signs to all distances measured along the positive direction and, conversely, negative signs to distances measured in the opposite direction.

In this connection, it is to be mentioned that a distance is measured *from* a certain reference point which may depend on the context, *to* some other point under consideration. For instance, in fig. 10-14), the object distance u , defined as the distance *from* N *to* O is *negative*, and so is the image distance v , which is the distance from N to I. This makes the equation (10-5) a consistent one since both the sides of this equation turn out to be positive. On the other hand, a wrong assignment of signs (say, a positive sign for u and a negative one for v) might have brought in an inconsistency in this equation.

A set of consistent rules for the assignment of signs to various quantities occurring in the mathematical relations in ray optics is referred to as a *sign convention*. Depending on the context, one may choose any one of various possible sign conventions in preference to another. However, having adopted one such convention, one has to stick to the same convention from the beginning to the end of a calculation so as to avoid inconsistencies. In sec. 10.9, I will present in greater details the sign convention commonly adopted in ray optics.

In the case of a single refraction at a plane surface considered above, the medium in which the object point is situated is commonly referred to as the *first* medium, while the medium which the rays are refracted into is termed the *second* medium. The equa-

tion (10-5) for image formation for paraxial rays can then be expressed as

$$\frac{\text{image distance}}{\text{object distance}} = \text{refractive index of the second medium relative to first.} \quad (10-7)$$

In figures 10-14 and 10-15 above, the second medium has been assumed to be optically lighter compared to the first ($n_2 < n_1$), i.e. the refractive index of the second medium relative to the first ($n = \frac{n_2}{n_1}$) has been assumed to be less than unity. If, on the other hand, the second medium is denser than the first, then n will be larger than unity and the image distance will be larger than the object distance, i.e., in other words, the image point will be located further away from the refracting surface compared to the object point, as in fig. 10-16.

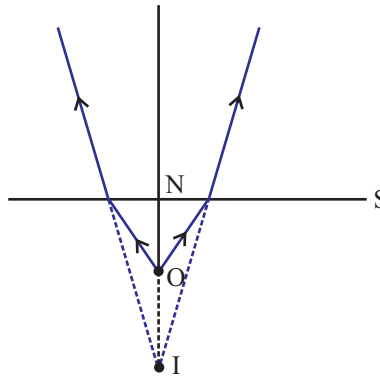


Figure 10-16: Refraction from a rarer to a denser medium - image formation by paraxial rays; the rays, on refraction, get bent toward the normal (ON); the image distance (NI) is larger in magnitude than the object distance (NO), in contrast to the situation depicted in fig. 10-14.

In optics, a medium is said to be 'denser' (resp. 'rarer') relative to another if its refractive index is greater (resp. smaller) than that of the latter. More specifically, the terms 'optically denser' and 'optically rarer' are used.

Problem 10-4

Imagine a horizontal layer of liquid of refractive index $n_1 = \frac{3}{2}$ and of thickness $D_1 = 0.04$ m, on which floats a horizontal layer of water of refractive index $n_2 = \frac{4}{3}$ and of thickness $D_2 = 0.02$ m. If

a point object located at the bottom of the first layer is viewed normally, where will its image be located?

Answer to Problem 10-4

Imagine a line normal to both the two liquid surface and passing through O, the object point. If N_1 and N_2 be the points of intersection of the line with the interface and the water surface respectively (draw your own figure; plot a set of ray paths for rays close to the line ON_1N_2 , and check the validity of the reasoning I give below) then $N_1O = -D_1$ and $N_2N_1 = -D_2$. Let I' denote the position of the image formed by the refraction at the interface between the two layers. Then, making use of eq. (10-5), one has, $\frac{n_2}{-v' + D_2} = \frac{n_1}{-D_1}$. In writing these equations, I have chosen N_2 as the reference point and the vertically upward direction as the reference direction so that, for instance, v' is the distance from N_2 to I' (check that the above equation correctly takes into account the sign convention I have referred to above).

I' now acts as the object for refraction at the water surface, giving rise to the final image, say, I , for which the distance v from N_2 satisfies $\frac{1}{v} = \frac{n_2}{v'}$. Combining the two, we get $v = -\frac{D_1}{n_1} - \frac{D_2}{n_2}$. Making use of given values for the present case, one gets $v = -0.042$ m, (approx), i.e., the final image is located at a point 0.022 m below the interface between the liquid and water layers.

The reasoning employed here holds in more general situations involving successive reflections and refractions in optical systems (see, for instance, sec. 10.12.1).

Applying similar considerations to the case of, say, k number of layers of thicknesses D_1, \dots, D_k (counting from bottom upward), and of refractive indices n_1, \dots, n_k , the final image may be seen to be situated at a distance $v = -\sum_{i=1}^k \frac{D_i}{n_i}$ from the top surface, where once again the same sign convention as the one indicated above is adopted.

10.5.2 Refraction through a layer with parallel surfaces

Fig. 10-17 depicts a layer of a transparent medium with parallel boundary surfaces A_1B_1 and A_2B_2 (sections by the plane of the figure), the refractive index of the medium being, say, n . Suppose that the refractive index of the medium on either side of the layer is n_0 . The figure shows a ray PQ incident on the first boundary surface at the point Q , where it is refracted into the layer along the path QR , being incident on the second surface at

R. The ray then undergoes a second refraction, emerging into the medium on the other side along RS.

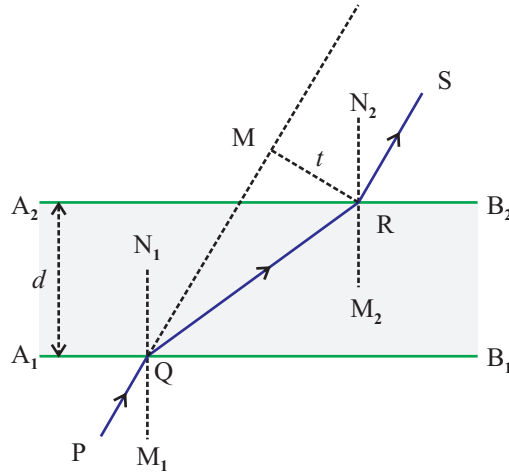


Figure 10-17: Refraction of a ray (PQ) through a layer with parallel boundary surfaces (A_1B_1 and A_2B_2); the ray emerges along RS, suffering a lateral deviation (t).

Problem 10-5

With reference to fig. 10-17, show that the emergent ray SR is parallel to the incident ray PQ, making use of the fact that the faces A_1B_1 and A_2B_2 are parallel to each other.

Answer to Problem 10-5

In the figure, N_1M_1 and N_2M_2 are perpendiculars to the two faces of the layer at Q and R respectively. Evidently, $\angle N_1QR = \angle M_2RQ$. Now apply Snell's law for refraction at Q and R to show that $\angle M_1QP = \angle N_2RS$.

According to the result obtained in exercise 10-5, the incident ray PQ and the emergent ray RS are parallel to each other. The distance t ($=RM$, where M is the foot of the perpendicular dropped from R onto the extension of the incident ray path PQ) is termed the *lateral displacement* of the ray in traversing the layer under consideration.

Problem 10-6

Referring to problem 10-5 and to fig. 10-17, obtain an expression for the lateral displacement of the ray in terms of the thickness d of the layer, the refractive indices n and n_0 , and the angle of incidence $i = \angle PQM_1$.

Answer to Problem 10-6

From the geometry of the figure, $t = QR \sin(r - i) = d \sec r \sin(r - i)$, where $r = \angle N_1QR$. Now apply Snell's law to obtain

$$t = d \sin i \left(\frac{n_0 \cos i}{\sqrt{n^2 - n_0^2 \sin^2 i}} - 1 \right). \quad (10-8)$$

10.6 Total internal reflection

Figure 10-18 shows three rays P_1N , P_2N , and P_3N incident at the point N on the surface S separating two media of refractive indices n_1 and n_2 , where $n_1 > n_2$, the former being the medium containing the incident rays. Of the three rays, the angle of incidence (θ) is the least for the ray P_1N , for which the reflected ray is NQ_1 and the refracted ray is NR , i.e., part of the energy carried by the incident ray P_1N along the direction normal to the interface (S) is sent back to the medium of incidence by means of the ray NQ_1 while the rest is transmitted on to the second medium by means of the ray NR .

Digression: components of energy flow.

There is something tricky here one needs to pay attention to. The flow of energy in the first medium (as also in the second medium) can be thought to be made up of two components - one perpendicular to the interface S , and the other parallel to it. Of the two, it is only the former that is involved in reflection while the parallel component (which exists in both of the media involved) does not undergo reflection or refraction. The normal component in the first medium can be seen to be the resultant of two parts, one associated with the incident ray and the other, propagating in the opposite direction compared to the former, with the reflected ray. The ratio of the reflected component to

the incident component of energy flux gives a quantitative measure (the *reflectivity*) of how efficient the surface S is as a reflector.

Meanwhile, the parallel component cannot, in general, be so decomposed, and one speaks instead of a single parallel component of energy flow in either of the two media. The normal component of energy flux in the second medium, when compared with the incident normal component in the first medium, gives the *transmissivity*.

The corresponding ratios, when referred to in terms of the respective *complex amplitudes*, are termed the *reflection coefficient* and *transmission coefficient*. Here I do not enter into an explanation of the term ‘complex amplitude’. A sinusoidally varying quantity, characterized by an amplitude and a phase, can be described in terms of a complex amplitude. You will find this discussed in some details in section 13.5.1.3.

The package of working principles referred to as ‘ray optics’ includes a set of rules from which one can work out the reflectivity and transmissivity for a given pair of media and for a given angle of incidence. The resulting equations are known as Fresnel Formulae.

The critical angle and beyond.

Since the second medium is rarer than the first, the angle of refraction is greater than the angle of incidence (see eq. (10-4b)) and, as the latter is made to increase, the former increases at a greater rate. The angle of incidence for the ray P_2N is such that the angle of refraction is $\frac{\pi}{2}$, i.e., the refracted ray grazes along the surface of separation between the two media. Evidently, this is the largest possible value for the angle of refraction, as a result of which, for a ray with a larger value of the angle of incidence (such as P_3N), there is no refracted ray by means of which energy is passed on to the second medium along the normal direction.

Instead, the entire normal component of energy flow associated with P_3N is reflected back into the first medium by means of the reflected ray NQ_3 . This is expressed by saying that the ray P_3N (as also any ray whose angle of incidence is greater than that

for the ray P_2N) undergoes *total internal reflection*.

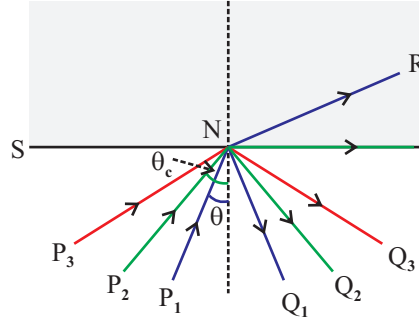


Figure 10-18: Illustrating total internal reflection; of the three incident rays, P_1N , P_2N , P_3N , the first ray undergoes reflection and refraction, while the second ray is refracted along the surface of separation; the third ray suffers total internal reflection; the surface S separates a denser medium (refractive index n_1) from a rarer medium (refractive index n_2), where the former is the medium of incidence.

The angle of incidence for the ray P_2N is given by

$$\theta_c = \arcsin \frac{n_2}{n_1}, \quad (10-9)$$

as can be checked from eq. (10-4a) by making use of the fact that the angle of refraction for this ray is $\frac{\pi}{2}$. One can thus say that, for a ray with its angle of incidence greater than θ_c given by the above expression, there occurs total internal reflection without a refracted ray being sent into the second medium.

A more complete understanding of the phenomenon of total internal reflection is obtained in *wave optics* where the propagation of light is interpreted as a transmission of electromagnetic disturbance from one region of space to another. The fraction of energy transmitted into the second medium along the normal direction can be worked out from electromagnetic theory and the result shows that, for $n_2 < n_1$, this fraction goes to zero precisely when the angle of incidence is θ_c given above.

For an angle of incidence $\theta > \theta_c$, the energy flow into the second medium in a direction perpendicular to the surface of separation continues to be zero, though there occurs an energy flow parallel to the surface. While ray optics gives the result that there is no

refracted ray transmitted into the second medium, wave optics tells us that there does remain a spatially- and temporally varying electromagnetic field in the latter without, however, any energy being transmitted along the normal direction. The electromagnetic field in the second medium is seen to be of a very special nature, and differs fundamentally from the field in the first medium.

The angle θ_c is referred to as the *critical angle* of the second medium with respect to the first. It may be mentioned in this connection that, regardless of whether the angle of incidence in the first medium is greater than or less than θ_c , the reflected ray in the first medium remains related to the incident ray in accordance with the laws of reflection.

Fig. 10-19 depicts a situation where the planar surface S separates two media marked '1' and '2', the former being optically denser with respect to the latter (relative refractive index $\frac{n_1}{n_2} = n(> 1)$). Rays fanning out from the point O in the medium '1' suffer total internal reflection unless the angle of incidence θ is less than θ_c given by (10-9). In other words, all rays emitted by a point source at O within a cone of semi-vertical angle θ_c (dotted lines) emerge into the medium '2', where the radius of the base of the cone is $NA = d$, with

$$d = h \tan \theta_c = \frac{h}{\sqrt{n^2 - 1}}, \quad (10-10a)$$

h being the distance of O from the surface S. Conversely, a ray incident on S from the medium '2' and reaching the point O must have its point of incidence P within the base of the cone, i.e., $NP = u$ (say) has to be less than d , where the angle of incidence i in medium '2' has to satisfy

$$\sin i = \frac{nu}{\sqrt{u^2 + h^2}} \quad (u \leq d). \quad (10-10b)$$

Problem 10-7

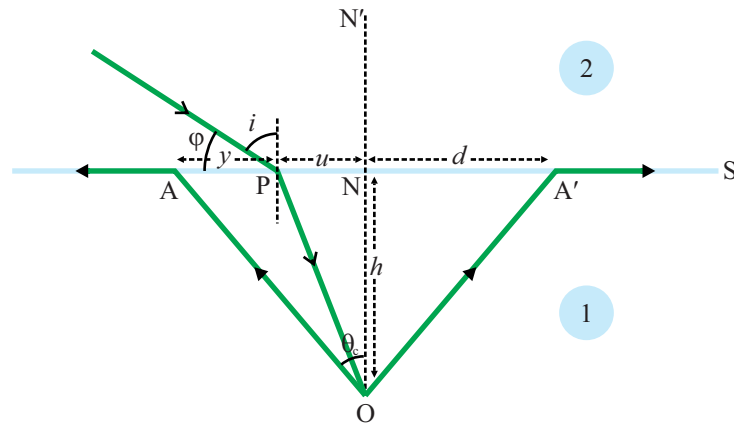


Figure 10-19: Limiting cone for refraction from an optically denser to a rarer medium, marked '1' and '2', separated by a planar surface S; rays emanating from the point O located in medium '1' emerge into medium '2' only if these are confined to the opening of a cone of semi-vertical angle θ_c ; AA' denotes the section of the base of the cone by the plane of the figure, while ONN' is the normal to S passing through O; the radius d of the base is given by formula (10-10a), where n stands for the refractive index of medium '1' with respect to '2', and h for the depth of O below S; a ray is shown incident from medium '1' and refracted into '2' along PO, where ϕ denotes the grazing angle of incidence (complementary to i , the angle of incidence); small values of ϕ correspond to low transmissivity; the expression for the distance AP ($= y$) can be worked out as in problem 10-7

Referring to fig. 10-19, consider a ray incident from medium '2' at a very small grazing angle ϕ (complementary to the angle of incidence i). Work out the expression for the distance AP in terms of h , n , and ϕ (up to the second degree in ϕ) if the refracted ray is to pass through O.

Answer to Problem 10-7

HINT: Making use of formulae (10-10a), and (10-10b), the required expression is obtained from $AP = d - u = \frac{h}{\sqrt{n^2 - 1}} - \frac{h \cos \phi}{\sqrt{n^2 - \cos^2 \phi}}$ where, up to the second degree in ϕ , $\cos \phi \approx 1 - \frac{\phi^2}{2}$. This gives $AP = y(\text{say}) \approx \frac{hn^2\phi^2}{2(n^2 - 1)^{\frac{3}{2}}}$.

NOTE: The transmissivity, i.e., the fraction of incident energy flux (in a direction perpendicular to the interface) entering into medium '1' from medium '2', which can be worked out from the relevant *Fresnel formulae* (refer to sec. 10.6) turns out to be small for small values of the grazing angle of incidence, going to zero for $\phi \rightarrow 0$.

10.7 Prism

Fig. 10-20 shows a *prism* - with a triangular base ABC and an edge AD, the latter being perpendicular to the base. In general, the term prism refers to a class of three-dimensional forms where the base need not be triangular - for instance it can even be circular, in which case it is termed a cylinder. But in optics, a prism means a transparent body with a triangular base with an edge perpendicular to the base. Moreover, in optics, the rectangular face BCFE opposite the edge, rather than the triangular face ABC, is often referred to as the base - a practice we will follow. In the glass prisms commonly used in laboratories, the base is usually made opaque to light rays.

The faces ABED and ACFD are termed the refracting surfaces of the prism because light rays are refracted, or bent, at these faces, following ray paths like PQRS shown in the figure. Here, the ray PQ is incident on the refracting face ABED and enters the material of the prism as QR, being refracted for a second time at the point R on the other refracting face ACFD, and emerging from the prism as ray RS.

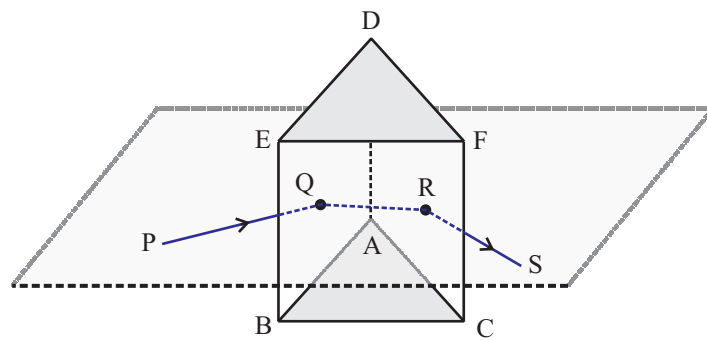


Figure 10-20: Bending of a ray due to two successive refractions at the refracting surfaces of a prism; the face BCFE is referred to as the base of the prism.

The ray path PQRS is contained in a plane cutting through the prism, perpendicular to the edge. It is relatively easy to describe paths of rays contained in planes such as this and we will consider only such rays in our discussion to follow. The bending of other rays, not contained in planes like the one shown in the figure, while more difficult to describe, however, occurs in accordance with the laws of refraction.

10.7.1 The basic formulae

Suppose that the refractive index of the material of the prism relative to that of the medium outside is n , which we assume to be greater than 1 (i.e., the material of the

prism is optically denser than the external medium). In general, the relative refractive index depends on the wavelength of light, which is why the prism is used for the ‘splitting of the colors’ of white light, but assume for the time being that the incident light is monochromatic so that, for a given incident ray, we will have only one refracted ray rather than a number of rays corresponding to different wavelengths.

One then has

$$\sin i_1 = n \sin r_1, \quad \sin i_2 = n \sin r_2, \quad (10-11)$$

(the symmetry of these two expressions is a result of our notation), and, from the geometry of fig. 10-21,

$$A = r_1 + r_2, \quad (10-12)$$

(check this out!).

One can, in principle, use (10-11) to eliminate r_1 and r_2 and to express the angle of emergence i_2 in terms of the angle of incidence i_1 for a prism of given angle A and given refractive index n . One can also work out the *deviation* in terms of i_1 , A , and n .

10.7.2 Deviation

Looking at fig. 10-21, the angle δ_1 between the path of the incident ray PQ and the refracted ray QR is the deviation caused by the refraction at the first refracting surface, while the angle δ_2 similarly depicts the deviation caused by the refraction at the second surface. The *total* deviation is the angle δ between the paths of the incident and the emergent rays, and it is evident from the geometry of the figure that

$$\delta = \delta_1 + \delta_2, \quad (10-13)$$

i.e., the two deviations $\delta_1 = i_1 - r_1$ and $\delta_2 = i_2 - r_2$ add up to give the total deviation δ as they should, since the both are in the same sense (clockwise, looked at from the top of

the figure). One thus obtains, on making use of (10-12),

$$\delta = i_1 + i_2 - A. \quad (10-14)$$

You can, if you like, express i_2 in terms of i_1 and A (the refractive index n will also make its appearance while you do so) in this formula, but the above form is neat and looks symmetrical in the two angles i_1 and i_2 .

Indeed, if you assume that the incident ray comes along SR (i.e., along the reversed emergent ray) then the bent ray path will be precisely SRQP, i.e., the ray will just retrace its previous path backwards. This is a result of the symmetry inherent in the laws of reflection and refraction according to which the incident and refracted (or reflected) rays are interchangeable.

For this reversed ray path the deviation has to remain the same (though it will now be in the opposite sense; but let us agree not to refer explicitly to the signs of angles, working instead with their magnitudes alone (see sec. 10.7.4)), which is precisely what the right hand side of (10-14) tells us since the reversed ray path just corresponds to an interchange of i_1 and i_2 .

Let us now do a bit of hard thinking and try to imagine what the graph of δ against i_1 would look like if we replace i_2 in (10-14) with its expression in terms of i_1 and A (not a simple-looking expression, I warn you, but we won't be needing this complicated formula, after all). What I can say outright is that, for each value of the deviation δ , there will be, in general, *two* different values of i_1 on the graph, i.e., the graph will be in the nature of the one shown in fig. 10-22.

To see why this should be so, consider an angle of incidence (i_1) for some particular value of deviation (δ), for which there corresponds some particular value of the angle of emergence (i_2). Now imagine a different value of the angle of incidence, say, i'_1 , where i'_1 is *this* value (i_2) of the angle of emergence. For this *new* angle of incidence, the value of angle of emergence will be i_1 , the angle of incidence we *started with* and so, from (10-14),

the deviation will *remain unaltered*. For instance, if in a given situation, $\delta = 50$ degrees for $i_1 = 40$ degrees and if, for this angle of incidence, the angle of emergence is, say 44 degrees then, with the angle of incidence $i'_1 = 44$ degrees, we will once again have $\delta = 50$ degrees.

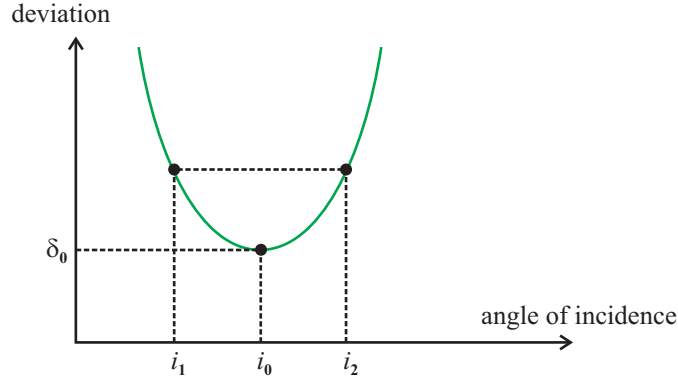


Figure 10-22: Graph of deviation against angle of incidence; in general, a given value of deviation corresponds to two possible angles of incidence; the two values coincide ($i_1 = i_2 = i_0$) at the lowest point of the graph, for which the deviation is minimum (δ_0).

10.7.3 Limiting angle of incidence

Looking back at fig. 10-21, note that I have assumed the incident ray to be confined within the region bounded by QN (the normal to GH at the point of incidence), and QH, and not within the region bounded by QN and QG. Likewise, the emergent ray is confined to the region bounded by RM and RI (and not within the region bounded by RM and RG). This ensures that the deviations δ_1 and δ_2 add up to produce the resultant deviation δ . If the angle of incidence i_1 is increased to such a value that the refracted ray QR within the prism is normal to the second face GI, then one has $r_2 = i_2 = 0$, and a larger value of i_1 would then imply that the emergent ray will lie within the region between RG and RM, and the condition for addition of the two deviations would be violated. For this limiting angle of incidence, we would have $r_1 = A$, and hence $\sin i_1 = n \sin A$. In other words, the limiting angle of incidence is given by

$$(i_1)_{\text{lim}} = \arcsin(n \sin A) \quad (\sin A \leq \frac{1}{n}), \quad (10-15a)$$

where the condition $\sin A \leq \frac{1}{n}$ implies that the limiting angle of incidence is less than $\frac{\pi}{2}$. If, on the other hand, the angle of the prism is larger than $A_{\text{lim}} \equiv \arcsin \frac{1}{n}$ then, since the maximum possible value of i_1 is $\frac{\pi}{2}$, all incident rays confined within the region bounded by QN and QH satisfy the condition for the addition of the two deviations:

$$(i_1)_{\text{lim}} = \frac{\pi}{2} \quad (\sin A > \frac{1}{n}). \quad (10-15b)$$

10.7.4 Minimum deviation

Now imagine a situation in which $i_1 = i_2$ (fig. 10-23), i.e., one in which the incident and the emergent rays are symmetrical with respect to the vertex A (I have renamed the section GHI of fig. 10-21 as ABC for the sake of convenience; this is not to be confused with the bottom face ABC of fig. 10-20). Evidently, the deviation δ corresponding to this situation must be such that the two values of the angle of incidence leading to this value of δ are the *same*. In other words, this special situation corresponds to the point P in fig. 10-22. Notice from the figure that this point on the graph gives us the *minimum* possible value of δ . We denote this special, minimum, value of deviation with the symbol δ_0 and call it the ‘angle of minimum deviation’ produced by the prism under consideration. The value of the angle of incidence (and also the angle of emergence) for this special situation we denote as i_0 . Evidently, from (10-14), one has

$$i_0 = \frac{A + \delta_0}{2}. \quad (10-16)$$

Since the two angles i_1 and i_2 are equal, r_1 and r_2 must also be so (see (10-11)), and calling this common value r_0 , equation (10-12) gives

$$r_0 = \frac{A}{2}, \quad (10-17)$$

and so, finally, equation (10-11) leads to the following relation between the angle of the

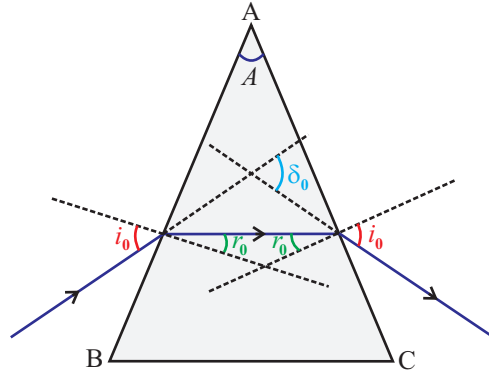


Figure 10-23: Ray path through a prism corresponding to minimum deviation; ABC is the section of the prism containing the ray path; minimum deviation corresponds to $i_1 = i_2 (= i_0)$ and $r_1 = r_2 (= r_0)$; the ray path is symmetrical with reference to the vertex and the refracting surfaces.

prism, the angle of minimum deviation, and the refractive index of the prism:

$$n = \frac{\sin \frac{A + \delta_0}{2}}{\sin \frac{A}{2}}. \quad (10-18)$$

It may be mentioned here that the above formulae pertaining to refraction of ray paths through a prism involve angles for which signs are not taken into consideration. In other words, all the angles are to be considered as positive, corresponding to situations depicted in the figures given above. In general, however, the formulae of geometrical optics hold between quantities bearing appropriate signs in accordance with the rules indicated in sec. 10.9.

Problem 10-8

The angle of incidence of a ray on the first refracting surface of a prism is $i_1 = \frac{\pi}{4}$. If the refractive index of the material of the prism be $n = \sqrt{2}$, and the angle of the prism be $A = \frac{5\pi}{12}$, show that the emergent ray grazes the second refracting surface, and obtain the deviation.

Answer to Problem 10-8

HINT: Considering the first refraction, the angle of refraction (r_1) satisfies $\sin r_1 = \frac{\sin i_1}{n}$, i.e., in the present instance, $r_1 = \frac{\pi}{6}$. Thus, the angle of incidence on the second surface is $r_2 = A - r_1 = \frac{\pi}{4}$. This happens to be the critical angle for the material of the prism, since the angle of emergence is seen to be $i_2 = \frac{\pi}{2}$. The deviation of the ray is $D = i_1 + i_2 - A = \frac{\pi}{3}$.

Problem 10-9

Referring to relations (10-11), (10-12), and (10-14), show that, for a *thin* prism, i.e., one for which the angle A is small (so that $\sin A \approx A$) the deviation is given by the simple formula

$$\delta \approx (n - 1)A. \quad (10-19)$$

Answer to Problem 10-9

HINT: For small angles the trigonometric sine functions can be replaced with the corresponding arguments. In this case, formula (10-15a) implies that the angle of incidence has to be less than nA for normal prism action (when the deviations produced at the two refracting surfaces of the prism add up to give the resultant deviation). Equations (10-11), (10-12), and (10-14) then imply $i_1 \approx nr_1$, $i_2 \approx nr_2$, i.e., $i_1 + i_2 = n(r_1 + r_2) = nA$, and hence $\delta \approx (n - 1)A$.

10.8 Reflection and refraction at spherical surfaces

10.8.1 Spherical mirrors: a few definitions

Figure 10-24(A), (B), (C) depict a plane mirror, a *concave* mirror, and a *convex* mirror respectively (sections by the plane of the figure). Recall that a mirror is a surface separating two media such that rays coming from one side of the surface (on the other side of the shading) are reflected back, with little of the incident energy being refracted into the other side (i.e., into the side of the shading).

The concave and the convex mirrors are parts of spherical surfaces where, in either case, C denotes the center of the sphere, while the point P shown in the figure for either mirror is referred to as the *pole* (see below). C is termed the *center of curvature* of the mirror, while the radius of the spherical surface, measured as the distance from the pole to the center of curvature, is referred to as the *radius of curvature*.

Notice that, for the concave mirror, the center of curvature is located on that side of the mirror from which the incident rays come in while, for the convex mirror, it is located on

the opposite side. If a piece of glass in the shape of a part of a spherical surface is given a coating of a mercury layer on the side opposite to the center of curvature, then it will act as a concave mirror while, if the coating is applied on the same side as the center of curvature then it will act as a convex mirror. In general, reflection takes place from any surface separating two media. However, a thin coating of mercury or of any other appropriate metallic layer increases the reflectivity, i.e., the fraction of incident energy sent back by reflection to the medium of incidence.

In fig. 10-24(B) and (C), P is a point on the surface of the mirror, i.e., CP is a radius of the spherical surface, such that the mirror is symmetrical about it. As mentioned above, P is termed the *pole* of the mirror. Referring to the circular arc shown in either figure, which is the section of the mirror by the plane of the figure, P is the mid-point of the arc. The straight line OP (imagined to be extended on both sides) is referred to as the *axis* of the mirror.

Though the mid-point of the arc representing the section of a spherical mirror is commonly defined to be the pole, in reality, any point on the spherical surface can be treated as the pole and the corresponding radius, extended both ways, can then be taken to be the axis. The pole and the axis are relevant for our purpose in that all ray paths we will consider below will be confined to a plane containing the axis, and all object points and (short) extended objects will also be assumed to be in that plane. Put differently, if one considers ray paths confined to a single plane then, with reference to these ray paths, the pole is a point on the surface of the mirror lying in this plane, and is commonly chosen to be the mid-point of the aperture of the reflecting surface. With the pole P chosen appropriately, and the axis along the line CP, the rules of geometrical optics can be made use of for conveniently constructing ray paths for incident rays (and object points) lying in this plane.

Fig. 10-25 shows a ray AN incident at the point N on a concave mirror (similar considerations apply to a convex mirror as well) with the following characteristic features: (a) the angle between the ray and the axis OP is small, and (b) the point of incidence N is close

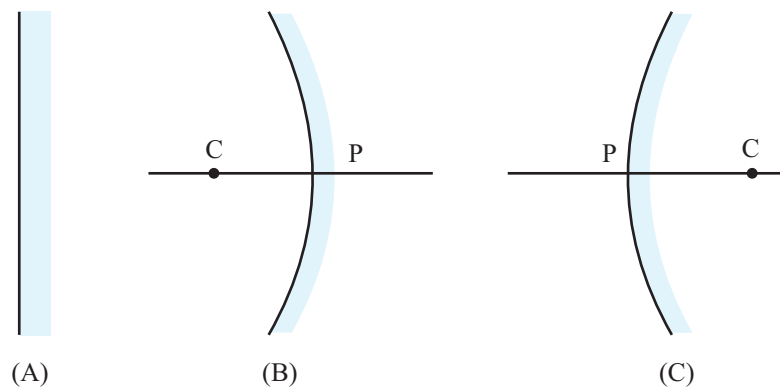


Figure 10-24: Plane and spherical reflecting surfaces (sections by the plane of the figure); (A) a plane mirror; (B) a concave mirror; (C) a convex mirror; in (B) and (C), the center of curvature C, and the pole P are shown (distances not to scale); the shadings show that the mirrors act as reflecting surfaces for incident rays coming from the left, with little of the incident energy being transmitted to the right.

to the pole P. Such a ray is termed a *paraxial* ray with reference to the mirror. In this book we will consider only such rays in the context of reflection and image formation by spherical mirrors. Non-paraxial rays, i.e., rays that do not satisfy the above conditions, introduce *aberrations* in the image of an object whereby the image fails to be a faithful replica of the object.

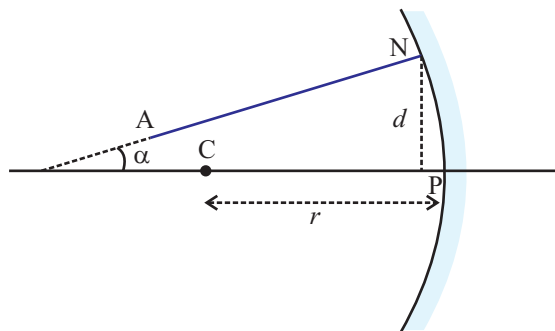


Figure 10-25: A paraxial ray AN incident at N on a concave mirror; the conditions for paraxiality are given by (10-20); similar conditions apply for reflection at a convex mirror as well.

1. There is nothing special about the sections shown in figure 10-24(B) or (C), any other plane section of the spherical surface under consideration containing the center of curvature being as good. For any such section, however, the pole P would be the same, being the mid-point of the corresponding circular arc. However, in

the context of reflection and image formation by paraxial rays even the mid-point of the section does not bear any special relevance, and any other point, say, P' (not shown) on the mirror could be taken as the pole. One would then have to consider only those rays whose points of incidence are close to P' and which make small angles with CP' , the corresponding axis. The surface of the mirror and its circular periphery would then not be symmetrical about the pole. Finally, even the periphery of the mirror need not be a circle since all one needs while considering image formation by paraxial rays is a small part of the spherical surface close to the pole.

2. In confining our considerations to paraxial rays, we do not consider rays which are *skew* to the axis. A ray is said to be skew to the axis if a plane cannot be found that contains both the axis and the ray path. If the entire ray path of a skew ray lies close to the axis, then it qualifies as a paraxial ray. Though a consideration of these skew paraxial rays requires a more sophisticated analysis compared to what I present below, they do not imply features in image formation distinct from those obtained with paraxial rays that are co-planar with the axis, provided we confine our attention exclusively to *axially symmetric* optical systems.

Here is a more specific characterization of a paraxial ray. Suppose that the distance of the point of incidence N (fig. 10-25) from the axis is d and the angle made by the ray with the axis is α (all angles will be expressed in radians). Then the condition for paraxiality can be expressed as

$$\alpha \ll 1, \quad d \ll r, \quad (10-20)$$

where r stands for the radius of curvature of the spherical mirror. These conditions can be made even more precise by stating that in our calculations relating to the ray paths we will ignore all terms involving the squares and higher powers of α and $\frac{d}{r}$. By implication one can, for instance, use the approximations $\sin i \approx i$ and $\tan i \approx i$, where i stands for the angle of incidence of the ray (see, for instance, fig. 10-26). However, in our pictorial representations of the ray paths, the figures will not be to scale and the conditions (10-20) may not be apparent in these figures.

In figure 10-26 (A) below, AN denotes a paraxial ray incident at the point N of a concave mirror, for which the reflected ray path is NQ. C being the center of curvature of the spherical surface of the mirror, CNC' is normal to the reflecting surface at the point of incidence N. Here $\angle ANC$ and $\angle CNQ$ are the *angles of incidence and reflection* respectively:

$$\angle ANC = i, \angle QNC = i'. \quad (10-21a)$$

Fig. 10-26(B) depicts a convex mirror on which the ray AN incident at N similarly gives rise to the reflected ray NQ, but now the angles of incidence and reflection are respectively

$$\angle ANC' = i, \angle QNC' = r, \quad (10-21b)$$

where C' is any point on the extension of CN.

The two angles i and i' satisfy

$$i = i', \quad (10-21c)$$

which expresses the law of reflection (eq. (10-3)) in the present context.

I end this section by referring to the earlier note on *sign convention* in ray optics in section 10.5.1. Indeed all the angles and distances in equations (10-20) and (10-21a), (10-21b), (10-21c) are, in reality, *signed* quantities, i.e., should carry appropriate signs (+ or -), which I have ignored temporarily. In other words, I have replaced all these signed quantities with their positive magnitudes. I will dwell in greater detail on sign convention in a subsequent section (sec. 10.9).

10.8.2 Refraction at a spherical surface: definitions

Figure 10-27 depicts a section of (A) a concave and (B) a convex spherical refracting surface separating two media of refractive indices n_1 and n_2 . Following the characteri-

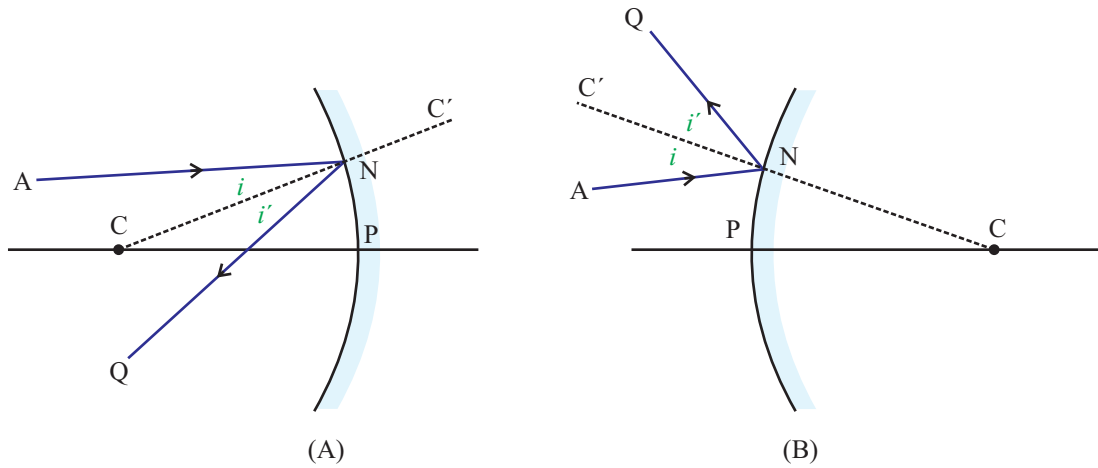


Figure 10-26: Reflection of a paraxial ray AN incident at the point N on a spherical mirror (schematic); (A) a concave mirror, (B) a convex mirror; the angles of incidence and reflection (resp., i and i') are shown.

ization of concave and convex mirrors, recall that the surface is called concave or convex according to whether or not the centre of curvature (C) lies on the same side as the one from which light is incident on it. We assume once again that the surface under consideration is part of the surface of a sphere, and consider only paraxial rays close to the axis CP, the definitions of centre of curvature, pole (P), and axis being as in the case of spherical reflecting surfaces (the figure shows an *axial* section in the sense that it contains the axis).

The radius of curvature (r) of the refracting surface is also defined similarly as the distance from the pole P to the centre of curvature C. The figure shows a ray AN incident on the surface at N from the first medium (refractive index n_1) and refracted into the second medium (refractive index n_2) along NQ. The line CNC' is normal to the refracting surface in this case, and the *angles of incidence and refraction* (respectively, i and r) are the ones shown in the figure. Thus, in fig. 10-27(A),

$$\angle ANC = i, \angle QNC' = r, \quad (10-22a)$$

while in fig. 10-27(B),

$$\angle ANC' = i, \angle QNC = r, \quad (10-22b)$$

and in both these cases,

$$n_1 \sin i = n_2 \sin r. \quad (10-22c)$$

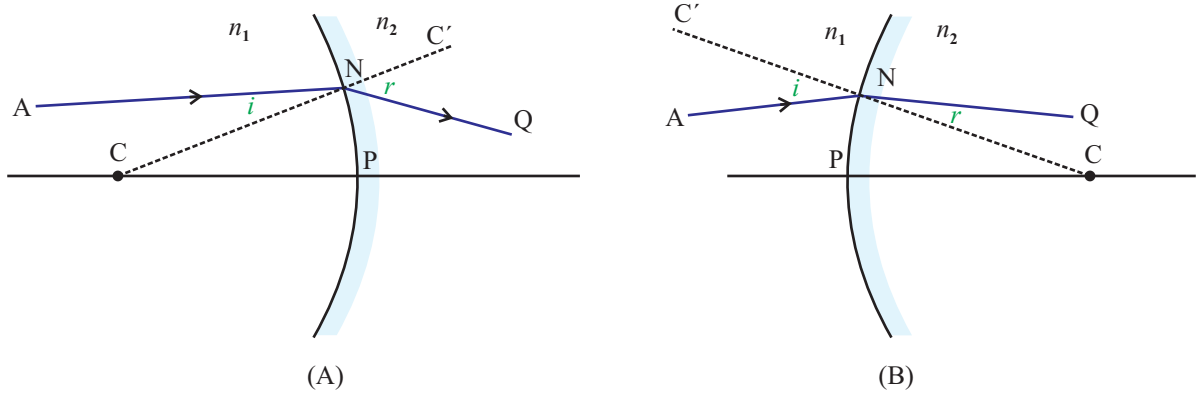


Figure 10-27: Refraction at a spherical surface separating two media; (A) a concave surface, and (B) a convex surface; the centre of curvature, pole, and axis are shown; a ray AN incident from the medium with refractive index n_1 is refracted along NQ; the angles of incidence and refraction (resp. i and r) are shown; the shading is used to distinguish between the media on the two sides of the surface, and not to indicate the reflecting surface of a mirror.

Once again, all the quantities introduced here are, in reality, to bear appropriate signs which I am temporarily ignoring (i.e., replacing the signed quantities with their positive magnitudes) till, in the next section, I tell you more on the sign convention to be followed in this book. The last relation (eq. (10-22c)) is a consequence of the laws of refraction in the present context.

Finally, as in the case of reflection in spherical mirrors, we will consider below the refraction of *paraxial rays* alone, because these are the rays that lead to image formation without aberrations. However, *skew rays* will not be considered even when the ray paths remain close to the axis since these require a more elaborate analysis as compared to

the *meridional* rays, a meridional ray being one for which there exists a plane containing both the ray path and the axis (the consideration of skew rays, however, does not lead to any new results in image formation since, in this book, we confine our attention exclusively to *axially symmetric* optical systems). The conditions for a ray to be paraxial are analogous to the ones mentioned in 10-20. Assuming that these conditions are satisfied for the ray AN in fig. 10-27, the relation (10-22c) can be written as

$$n_1 i = n_2 r, \quad (10-22d)$$

where, once again I have, for the time being, ignored the signs of the quantities involved.

10.9 Sign convention in ray optics

As I have mentioned before on several occasions (e.g., in sec. 10.5.1), one needs a *sign convention* in ray optics where the various distances and angles (as also *refractive indices*, see below) are to be assigned appropriate signs in order that the mathematical relations one uses or arrives at do not lead to inconsistencies.

The sign convention in respect of distances along the axis of a spherical reflecting or refracting surface can be stated as follows, where it may be recalled that a distance is measured *from* a certain reference point *to* some specified point or other:

If the axis of the spherical surface is assumed to be along the horizontal direction, then a distance measured from the left to the right will be taken as positive while one measured from the right to the left will be assumed negative.

To put it differently, imagine the x-axis of a Cartesian co-ordinate system to lie along the axis of the spherical surface under consideration, with the x-y plane being the one containing the ray paths. Let the positive direction of the x-axis be chosen from the left to the right in the ray diagram, where the diagram is drawn with the axis of the spherical surface extended along a horizontal line. Distances along the axis of the spherical surface are then taken to be positive if they are measured along the positive

direction of the x-axis.

Thus, for instance, in fig. 10-24(B), (C), the radius of curvature r , defined as the distance *from* the pole *to* the centre of curvature, is negative for a concave mirror and positive for a convex one. A similar statement applies to the concave and convex refracting surfaces in fig. 10-27.

For a distance measured perpendicular to the axis (referred to as a lateral or transverse distance), a plane may be imagined containing the axis and the line along which the distance lies (the x-y plane mentioned above), and a reference direction along the line (i.e., one parallel to the y-axis) is to be chosen as the positive direction. The distance is measured *from* the axis either in the positive or in the negative direction, corresponding to which the sign of that distance will be positive or negative. For instance, in fig. 10-25 or in a similar figure, distances measured from the axis *upward* are taken to be positive while those measured downward are assigned a negative sign. Thus, the distance d in 10-25, which points upward when measured from the axis is to bear a positive sign while the signs of other distances can be similarly assigned.

Note in this context that the sign convention for axial and transverse distances are simply the commonly accepted conventions one follows in co-ordinate geometry, where the x-axis corresponds to longitudinal distances and the y-axis to transverse ones.

I now consider the sign convention for *angles*. In order to see whether the angle made by a ray with the axis or with any other reference line is positive or negative, one has to find out the direction in which the reference line is to be rotated so as to coincide with the ray path. However, there are *two* distinct ways in which a reference line can be rotated to the position of the line under consideration, one being through a smaller (acute) angle and the other through a larger (obtuse) angle. Of these, the former is to be considered for assigning the sign to the angle made by the ray path with the reference line. If the sense of rotation is anticlockwise (looking from any chosen side of the plane containing the two lines under consideration), then the angle is taken to be positive while, in the case of a clockwise rotation, the angle is assumed negative.

Referring, for instance, to fig. 10-25, the angle α made by the ray AN with the axis CP is seen to be positive according to the above convention, where we choose to look at the plane of the figure from above, because the axis is to be rotated in an anticlockwise sense (when viewed from above) so as to make it coincide with the line AN. It may be noted in this connection that the direction of the ray is not pertinent in determining the sign of the angle - the sign would be positive even if the direction of the ray were from N to A (instead of being from A to N).

Finally, with reference to any given ray, one also needs to assign an appropriate sign to the *refractive index* (say, n) of a medium, depending on the direction of propagation of the ray. For a ray propagating from the left to the right (i.e., towards the positive direction of the x-axis), the refractive index will be assumed positive while the sign will be assumed negative for a ray propagating from the right to the left.

In order to see how the sign conventions for angles and refractive indices operate, look at fig. 10-28(A), (B), (C), (D), where four rays are shown. Denoting the angle made by any of these rays with the reference line OO' by α , one finds that α is positive for fig. 10-28(A) and (B), while in fig. 10-28(C) and (D) the sign of α is negative. On the other hand, the refractive index n is seen to be positive for fig. 10-28(A), (C), and negative for fig. 10-28(B), (D).

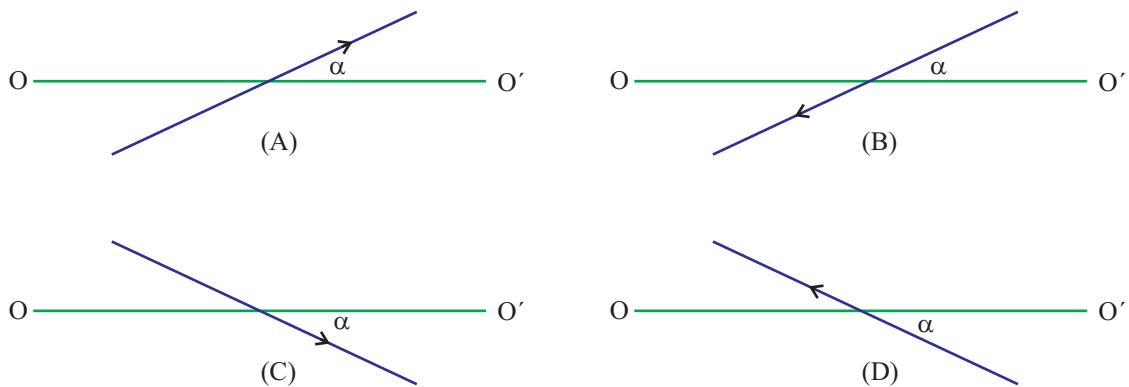


Figure 10-28: Illustrating the sign conventions for angles and refractive indices; the angle α between the ray and the reference line OO' is positive in (A) and (B), and negative in (C) and (D); on the other hand the refractive index n of the medium with reference to the ray is positive for (A) and (C), and negative for (B) and (D).

When one talks of the refractive index of a *medium* for a given *frequency* of light, it is commonly referred to as a positive quantity. There exists a theory, based on fundamental principles, from which one can calculate the refractive index, at least in a number of simple situations. However, in the context of a ray propagating in the medium, one needs to take into consideration the appropriate sign of the refractive index in order that inconsistencies may be avoided in calculations relating to ray optics.

In this book, we will not be concerned with the signs of angles and refractive indices, using only their magnitudes in the mathematical formulas we write down. The calculations in this book will be of a relatively simple nature, which will prevent inconsistencies coming in provided we exercise a bit of care in writing out the formulas. While referring to *distances*, however, the sign convention outlined above will be followed.

Nevertheless, I want to present you with an example of how a consideration of the signs of angles and refractive indices can modify the mathematical relations one arrives at in ray optics. Think, for instance, of the reflection of a ray from a surface, say, a concave one as in fig. 10-26(A), and of eq. (10-21c). For the angles i and i' , the normal CNC' is taken to be the reference line. Noting that this line is to be rotated in a clockwise direction so as to make it coincide with the incident ray path AN (extended both ways), the sign of the angle of incidence i is to be chosen negative. On the other hand, the same rule tells us that the sign of the angle of reflection i' is positive. Consequently, the correct relation between i and i' turns out to be

$$i = -i', \quad (10-23)$$

instead of eq. (10-21c).

In this context, it is of some interest to note that the equation (10-23) expressing the law of reflection can be looked upon as a special case of eq. (10-22c) expressing Snell's law of refraction (this equation does not get modified when one takes into account the signs of the quantities involved). This can be done by replacing the angle of refraction r in eq. (10-22c) with the angle of reflection i' , and the refractive index n_2 with the

refractive index of the first medium where the ray is sent back in reflection. However, since the ray now propagates in the opposite direction as compared to the incident ray, one has to replace n_2 with $-n_1$. With these substitutions, one does indeed find that eq. (10-22c) leads to eq. (10-23) which, rather than eq. (10-21c), is the correct equation to use for reflection. However, as mentioned above we will not concern ourselves with a consideration of signs of angles and refractive indices in our calculations.

10.10 Image formation in reflection by a spherical mirror

10.10.1 Focal length of a spherical mirror

Figure 10-29(A), (B) depict a concave and a convex mirror respectively, in each of which a paraxial ray AN parallel to the axis of the mirror is incident at N on the mirror, and gets reflected along NQ. The ray path (produced backward in (B)) intersects the axis at F. As we will see below, the location of F on the axis does not depend on the point of incidence N, i.e., the reflected ray paths (or their extensions backward) of *all* paraxial rays parallel to the axis pass through the *same* point F on the axis, which is referred to as the *focal point* (or, in brief, the *focus*) of the mirror.

One observes from the figure,

$$\angle NCF = \angle CNF, \text{ i.e., } |CF| = |FN|, \quad (10-24)$$

where $|CF|$ etc. denote the lengths of the respective line segments, which are all positive magnitudes.

Let us now recall the condition for the ray AN to be paraxial, namely, the point N should be sufficiently close to the pole P (this is not apparent from the figure, which is not to scale). If this be the case, then we can write, to a good degree of approximation, $|FN| = |FP|$, and one thus has, $|PF| = \frac{|CP|}{2}$. In other words, the point of intersection (F) of the reflected ray (or its extension backward) with the axis bisects the segment CP.

Since this holds for all paraxial rays parallel to the axis, we conclude that, for a bunch of paraxial rays all parallel to the axis, all the reflected rays converge to or appear to diverge from F , which we have termed above the focal point.

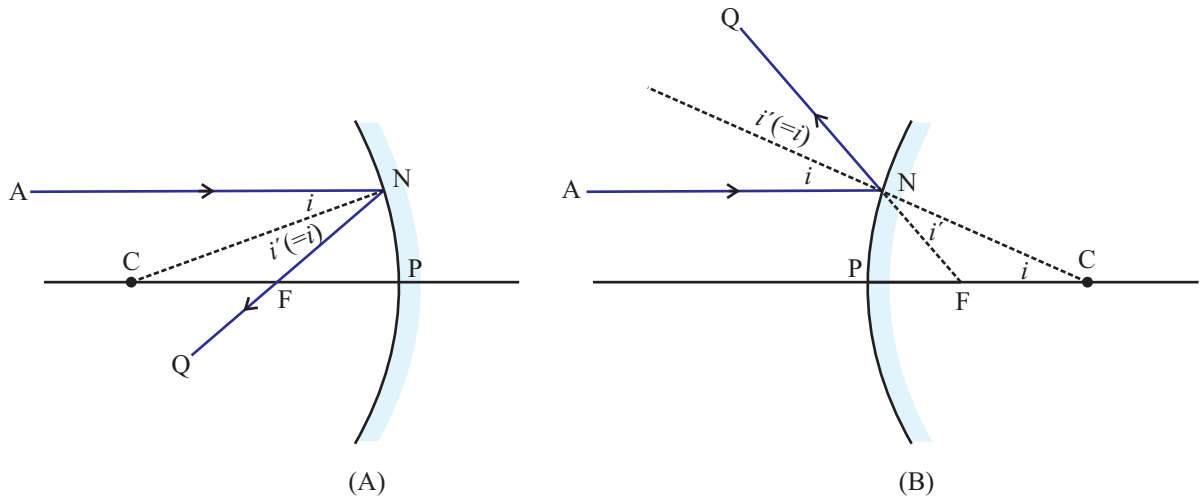


Figure 10-29: Paraxial ray parallel to the axis reflected from (A) a concave mirror, and (B) a convex mirror; the reflected ray path (or its extension backward in (B)) intersects the axis at F , the focal point of the mirror.

One has to keep in mind the relation of the *signed distances* like the radius of curvature (r) and focal length (f) with *lengths* of corresponding line segments in accordance with our sign convention. For instance, for a concave mirror (fig. 10-29(A)), the radius of curvature r is negative, and we have $|CP| = -r$ (on the other hand, CP will denote the distance *from* C *to* P , and one can write $CP = -r$ or, similarly, $PC = r$; a symbol like CP or AN may denote either a directed line-segment or a directed line, like a ray path, depending on the context).

Thus, in terms of the signed distances, $PC = 2PF$, i.e.,

$$r = 2f, \quad (10-25)$$

where f denotes the focal length. For a concave mirror (fig. 10-29(A)), r and f are both negative quantities, and the above relation tells us that the focal length is half the radius

of curvature in magnitude.

In the case of a convex mirror (fig. 10-29(B)), one will similarly have $|FP| = |CF|$ (check this out), and once again, $r = 2f$, the same relation that holds for a concave mirror, but with the difference that r and f are now both positive.

10.10.2 Aperture

In fig. 10-30 the plane of the circular boundary (NBN'D) of a concave mirror is perpendicular to the axis CP, and the mirror is axially symmetric about CP.

The axial symmetry of the mirror which this implies is not the same thing as the axial symmetry that characterizes image formation by paraxial rays - the latter does not require that the rim of the mirror be axially symmetric. In the image formation by paraxial rays, only a small region of the mirror around the pole is relevant and the geometrical relation between the incident and the reflected rays remains invariant under a rotation about the axis. In the present section the rim of the mirror is assumed to be axially symmetric for the sake of clarity of presentation.

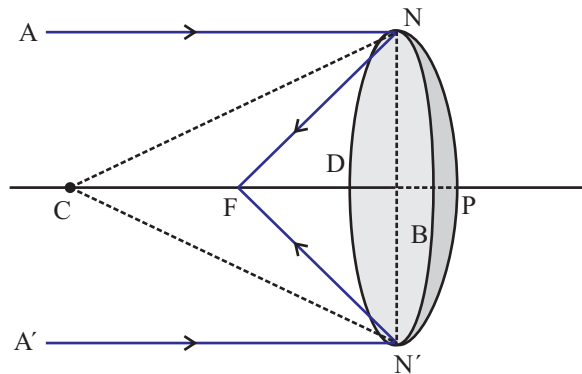


Figure 10-30: Illustrating the idea of aperture of a spherical mirror; the aperture is expressed quantitatively by the length of the diameter NN' or, equivalently, by the angle that this diameter subtends at the centre of curvature C ; the rim of the mirror is assumed to be axially symmetric about CP ; a mirror with a large aperture can accommodate a wide beam of rays within its extent, as a result of which there may be pronounced aberration in the images formed by it.

The ray AN , parallel to the axis, is incident at the point N of the rim, i.e., is situated at

the maximum distance from the axis that the physical extent of the mirror allows. $A'N'$ is another ray, again parallel to the axis, incident at N' , the point diametrically opposite to N . Then the distance between AN and $A'N'$, i.e., the diameter of the circular rim, gives a measure of how wide a beam can be accommodated within the extent of the surface of the mirror, and is referred to as its *aperture*. At times, the aperture - or, angular aperture as it is called - is expressed by the angle that the diameter NN' (or any other diameter) of the circular rim subtends at the centre of curvature C of the mirror.

Evidently, if the aperture of the mirror be small, then *all* rays parallel to the axis - or making a small angle with the axis - that the mirror can accommodate satisfy the condition of being paraxial rays. In this case, there will be only negligible defect in the image of an object formed by the mirror. On the other hand, for a mirror whose aperture is not small, the rays reflected from points near the rim of the mirror cannot be said to be paraxial, and these rays cause the shape of the image to deviate considerably from that of the object.

While a mirror with a small aperture is desirable from the point of view of perfection in image formation, it suffers from the disadvantage of admitting only a narrow beam of rays from, say, a point object, as a result of which the brightness of the image is compromised. In other words, in order to have an image of adequate brightness, one has to use a mirror of not too small an aperture. The consequent defect in image formation can be remedied to some extent by appropriately *shaping* the mirror, which would then no longer be a part of a spherical surface. For instance, a *paraboloidal* reflector forms a perfect image of a distant point object even with non-paraxial rays. This fact is made use of in optical telescopes, radio-telescopes, and microwave receiving antennas.

10.10.3 Image formation: relation between object distance and image distance

Fig. 10-31 (A) shows an axial section of a concave spherical mirror with center of curvature C , pole P , and focus F , in which A represents a point object situated on the axis somewhere to the left of F , so that A and P lie on two different sides of F . AN is a paraxial

ray incident at N on the mirror and reflected along NQ, say, where the reflected ray path intersects the axis at, say, I. As we will see below, the reflected ray path for any *other* paraxial ray like, say, AN', will also intersect the axis at the *same* point I. In other words, a divergent beam of rays originating from A will converge, on being reflected from the mirror, at I on the axis, whereafter these reflected ray paths will diverge away from I. As mentioned earlier, a ray path corresponds to the path along which electromagnetic energy is transported. In other words, there occurs a concentration of radiant energy at the point I due to reflection at the spherical mirror, the point of origin of this energy being A. In these circumstances, I is termed the image of A (recall our earlier discussion of image formation in sec. 10.3) formed by the spherical mirror.

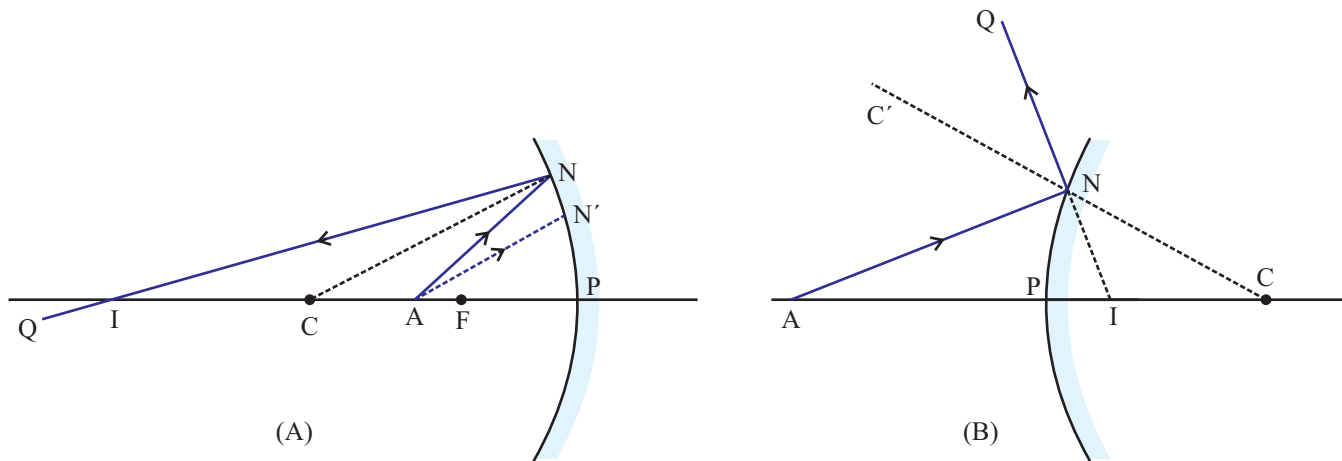


Figure 10-31: Image formation (schematic) of an on-axis object point by (A) a concave mirror, (B) a convex mirror; a paraxial ray originating from the point A located on the axis is incident at N and reflected along NQ; the latter (or its extension backwards) intersects the axis at I; other reflected rays (or their extensions backwards) also intersect the axis at the same point; I is then the image of the point object A; the image is real in (A), and virtual in (B).

Look at the triangles, $\triangle CNI$ and $\triangle ANC$ in fig. 10-31(A). According to the law of reflection, $\angle ANC = \angle CNI = i$ (angle of incidence). Then,

$$\frac{|IN|}{|CI|} = \frac{|AN|}{|CA|} = \frac{\sin \theta}{\sin i}, \quad (10-26)$$

where θ stands for the angle $\angle NCA$ (check this relation out!). We now make use of the

paraxiality condition, by virtue of which we can write (to a good degree of approximation)

$$|IN| = |IP|, \quad |AN| = |AP|. \quad (10-27)$$

We can then write, from equation (10-26)

$$\frac{|IP|}{|CI|} = \frac{|AP|}{|CA|}, \quad (10-28)$$

or, in other words,

$$\frac{|IP|}{|IP| - |CP|} = \frac{|AP|}{|CP| - |AP|}. \quad (10-29)$$

We define the *object*- and *image* distances as (these are distances with appropriate signs)

$$u = PA(= -|AP|), \quad v = PI = (-|IP|), \quad (10-30)$$

where I have indicated how these are related to the lengths of the respective line segments in fig. 10-31 (A). Notice that the object distance is, by definition, the signed distance *from* P *to* A, while the image distance is also similarly defined. Recalling that $r = -|CP|$, one then has, for the situation depicted in fig. 10-31 (A),

$$\frac{-v}{-v - (-r)} = \frac{-u}{-r - (-u)}, \quad (10-31)$$

which gives, on simplification,

$$\frac{1}{v} + \frac{1}{u} = \frac{2}{r}, \quad (10-32)$$

or, in view of (10-25),

$$\frac{1}{v} + \frac{1}{u} = \frac{1}{f}. \quad (10-33)$$

The great thing to note about this formula is that, while the point I has been defined with reference to the ray AN, its location, as obtained from (10-33), depends *only* on

the location of A on the axis, and *not* on the point of incidence N on the mirror. Thus, *all* rays diverging from A will, after reflection from the mirror, pass through the *same* point I on the axis, which can thereby be termed the image point corresponding to the object point A. Formula (10-33) then gives the relation between the object distance and the image distance (both measured from the pole P) for the spherical mirror, subject to the condition of paraxiality. Recall that, in this formula, u , v , and f are all distances with signs. While it has been derived for the situation depicted in fig. 10-31(A), it holds for other situations as well, where an image is formed for an on-axis object point by paraxial rays reflected from a spherical mirror.

For instance, fig. 10-31 (B) depicts image formation by a *convex* mirror. In this figure, I is the *virtual* image for the object point A since rays diverging from A appear to come from I after reflection from the mirror. For this, situation, u is a negative quantity, while v , r (and f) all positive. Nevertheless, the relation (10-33) continues to hold, provided only that the condition of paraxiality is obeyed.

One other feature of formula (10-33) is that, it is *symmetrical* with respect to u and v . This implies that if a point object were located at I, then its image would have been formed at A: object- and image points are interchangeable. This is part of a more general principle that tells us that ray paths are *reversible*. The interchangeability of object and image points is sometimes expressed by saying that they are *conjugate* to each other.

Finally, note the consistency of formula (10-33) with the definition of the focal point. The latter implies that the focus is the image point corresponding to an object point located at an infinite distance from the mirror i.e., one for which $u \rightarrow -\infty$. For such an object point, formula (10-33) implies that $v \rightarrow f$, i.e., the image will be formed at a distance f from the pole, corresponding precisely to the focal point.

Problem 10-10

A short object of length $l = 0.01$ m lies along the axis of a spherical mirror of focal length $f = -0.2$ m at a distance $u = -0.6$ m from it. What will be length of its image, and where will it be located?

Answer to Problem 10-10

HINT: Let the end points of the object be at distances u and $u + l$ from the mirror on the axis. If the corresponding end points of the image, also on the axis, be at distances v and $v + l'$, then one gets, on applying formula (10-33) to the two end points of the object in succession,

$$\frac{1}{v} + \frac{1}{u} = \frac{1}{f}, \quad \frac{1}{v + l'} + \frac{1}{u + l} = \frac{1}{f}.$$

Making use of these two relations and of the fact that l (and l') are small quantities, one obtains $m_1 = \frac{l'}{l} = -\frac{v^2}{u^2}$ (formula (10-35)). With the given values in the present problem, one obtains $v = -0.3$ m, and $m_1 = -\frac{1}{4}$, and hence $l' = -0.0025$ m. Note that the image and the object, both lying on the axis, are oppositely oriented with reference to the mirror, a feature common to all instances of image formation by reflection.

10.10.3.1 Image for an off-axis object point

Figure 10-32(A) depicts a point object A situated *off* the axis CP of a concave mirror, the height h above the axis being small compared to the radius of curvature (r) of the mirror (refer to the condition of paraxiality, eq. 10-20 where, in the present context, d is to be replaced with h). Imagine a bunch of paraxial rays diverging from A and incident on the mirror.

One can make use of the laws of reflection, as in the last section, to show that all these rays, after reflection from the mirror, pass through a certain point I, again located off the axis, which can thus be termed the image of A. In the figure, two of the ray paths from A to I are shown for the sake of illustration. One of these (ANFI) corresponds to an incident ray parallel to the axis, for which the reflected ray necessarily passes through the focus F while the other is incident normally on the mirror and retraces its path along MACI, so that the two intersect in I. It is interesting to note that AM need *not* be a paraxial ray for the derivation (which you will work out for yourself presently) to hold.

In the figure, AB and IJ are perpendicular to the axis, dropped from the object point A and the image point I respectively. PB and PJ - distances of A and I from P measured

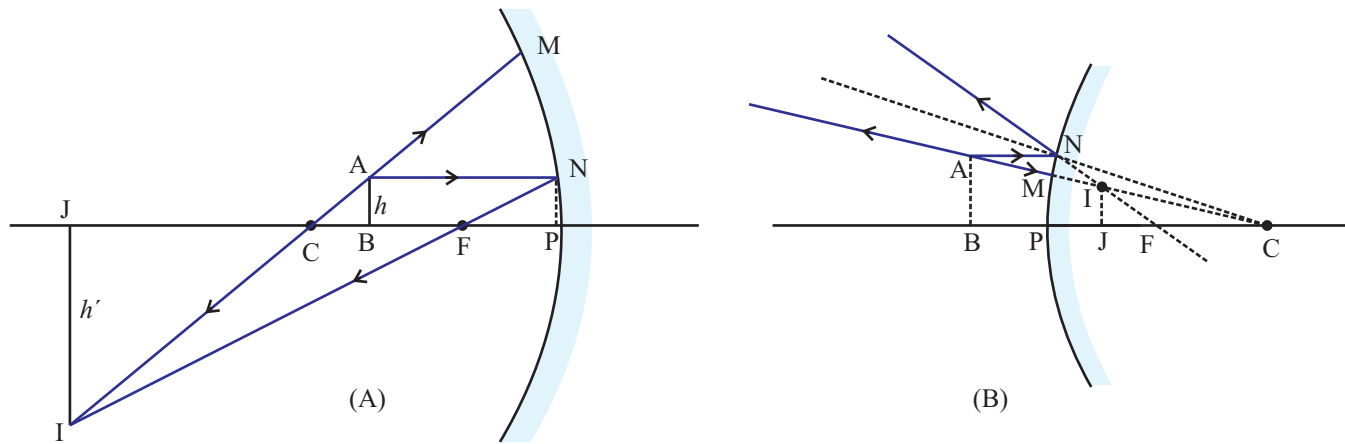


Figure 10-32: Image formation (schematic) for an off-axis object point by (A) a concave mirror, (B) a convex mirror; a bunch of rays diverging from A (two such rays are shown in each of (A) and (B)), after reflection from the mirror, converge to the point I in (A), and appear to diverge from I in (B); I is thus the image of A, real in (A) and virtual in (B).

along the axis are termed the object- and image distances respectively. Denoting these by u and v , one notes that u and v are both negative in figure 10-32(A), while the heights h and h' above the axis are respectively positive and negative. If O (not shown in the figure) be the foot of the perpendicular dropped from N onto the axis, the condition for paraxiality is that $|PO|$ is to be small (compared to u and v). In the following, we assume the point O to be coincident with the pole P.

I leave it to you to make use of the figure and to establish that, once again, u and v are related by (10-33). And once again, object- and image points are conjugate to each other (triangles ABC and IJC are similar, and so are triangles NPF and IJF; warning: take good care of the signs).

While you arrive at relation (10-33) for the situation depicted in fig. 10-32(A), the formula holds for other situations as well, where appropriate signs of u , v and f are taken into account. For instance, fig. 10-32(B) depicts a case of image formation by a convex mirror, where u is negative while v and f are positive, the image in this case being a virtual one.

10.10.3.2 Image formation for short extended objects

One other formula I should like you to work out in the course of the above derivation is:

$$\frac{h'}{h} = -\frac{v}{u}, \quad (10-34)$$

(look at triangles APB and IPJ; Make use of the fact that, for the incident ray AP, the reflected ray path passes through I, the image of A).

What is interesting here is that the ratio $\frac{h'}{h}$ depends only on the horizontal distances (measured from the pole) of A and I, and not on the vertical distance h of the object point from the axis. Hence, considering a series of object points, all lying on AB, all the corresponding image points will be formed on IJ (as implied by (10-33)), and for each object-image pair formula (10-34) will hold. In other words, if one places a short *extended* linear object AB perpendicularly to the axis along AB, an extended linear image IJ will be formed as in fig. 10-32 (A) or (B). What is more, all parts of the object will be magnified by the *same* ratio, namely, $-\frac{v}{u}$. This is known as the *transverse magnification* produced by the mirror.

Evidently, this transverse magnification (m_t) is the same for all linear objects placed in a plane perpendicular to the axis, all such objects being necessarily at the same distance from the mirror. In other words, a *planar* object placed perpendicularly to the axis will produce a *similar* planar image with the same magnification (fig. 10-33 (A)).

Finally, I leave it to you as an interesting exercise to define a *longitudinal magnification* (m_l) and to show that it is given by the formula

$$m_l = -\frac{v^2}{u^2}. \quad (10-35)$$

This implies that a *three-dimensional* object will *not* give rise to a similar three-dimensional image since the magnification along the axis will differ from that transverse to it (fig. 10-33 (B)).

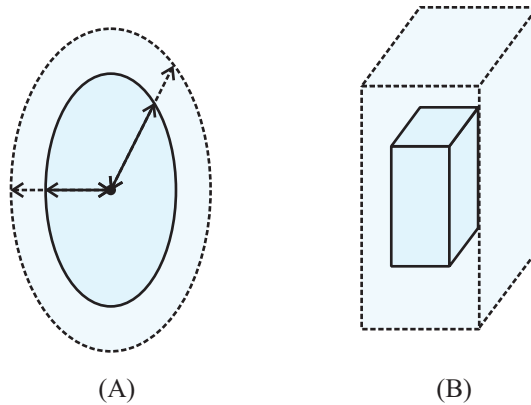


Figure 10-33: Illustrating the magnification (i.e., the ratio of the image size and object size for (A) a planar object and (B) a three-dimensional object; the image (dotted line) is placed over the object (solid line) for comparison; all line elements in the object are similarly magnified in the image in a transverse plane, but the magnification in a longitudinal direction occurs by a different ratio.

10.11 Image formation by refraction at a spherical surface

Fig. 10-34 illustrates the phenomenon of refraction at a spherical surface, where (A) and (B) depict refraction at a concave and a convex surface respectively.

A paraxial ray AN incident on the refracting surface in a direction parallel to the axis in the first medium (refractive index n_1) at the point N is refracted along NQ into the second medium (refractive index n_2). We assume for the time being that $n_2 > n_1$, i.e., the ray is refracted from an optically rarer to a denser medium. Then, in fig. 10-34(A), the ray path NQ , when produced backwards, intersects the axis at the point F situated in the first medium, while in fig. 10-34(B), NQ intersects the axis at F in the second medium. This point (F) is termed the *focal point* of the refracting surface.

For the sake of completeness, refer to fig. 10-35 (A) in which an incident ray AN gives a refracted ray NQ parallel to the axis, where the path AN produced on the other side of the concave refracting surface intersects the axis at the point F' . Similarly, in fig. 10-35 (B), a ray originating from the point F' on the axis and incident at N on a convex refracting surface is refracted along the path NQ , parallel to the axis. In either case,

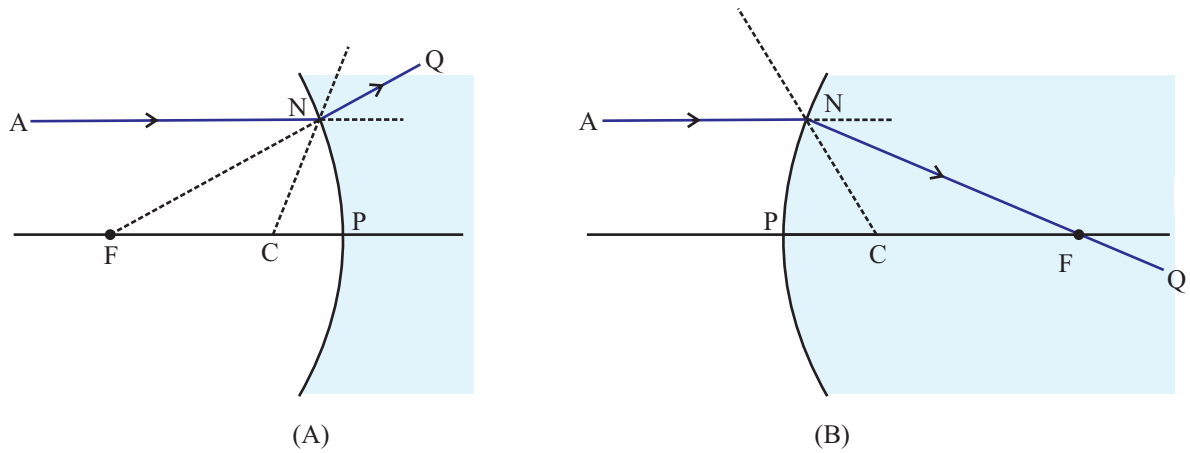


Figure 10-34: Refraction at a spherical surface : (A) concave surface, and (B) convex surface; a paraxial ray AN parallel to the axis CP is refracted along NQ; F is the focal point of the refracting surface.

the point F' is termed the *first* focal point of the refracting surface (we have once again assumed $n_2 > n_1$ while drawing the ray paths in fig. 10-35 (A), (B)). By contrast, the point F in fig. 10-34 is referred to as the *second* focal point. It is this second focal point that is usually termed the focal point (or, simply *focus*) in brief.

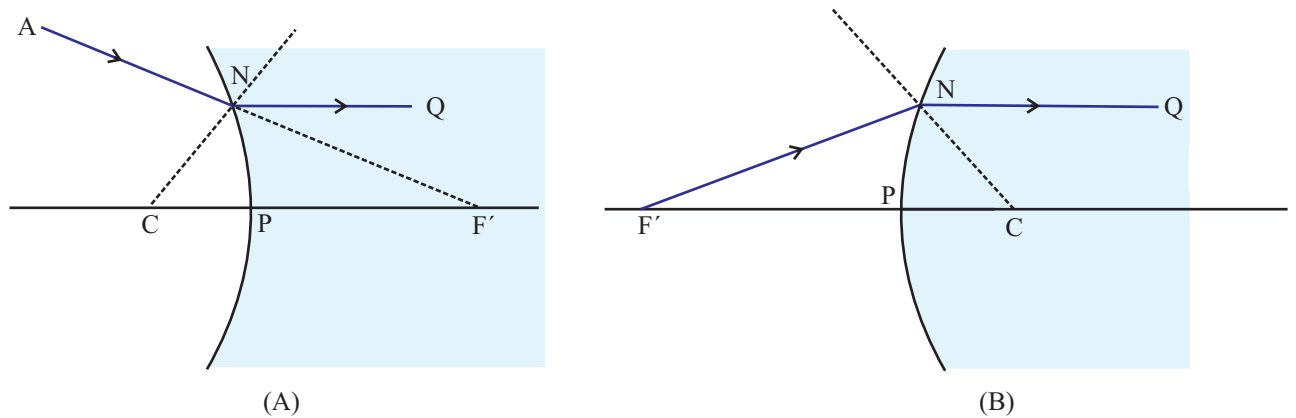


Figure 10-35: Defining the first focal point for refraction at a spherical surface: (A) concave surface, and (B) convex surface.

The definitions of the first and second focal points given here hold (with necessary but incidental modifications) even when $n_1 > n_2$. Moreover, the point F or F' turns out to be independent of the point of incidence N of the ray AN, provided the paraxiality condition

is satisfied. This latter fact implies that F is the *image* point for an object point located at an infinite distance, while F' is the object point for which the image is formed at an infinite distance.

10.11.1 Refraction at a spherical surface: image formation for a point object on the axis

Fig. 10-36 shows a point object A on the axis (PC) of a convex refracting surface with centre of curvature C and pole P . A paraxial ray AN originating from A in the first medium (refractive index n_1) is refracted at N , giving rise to the refracted ray NQ in the second medium (refractive index n_2), where the ray path intersects the axis at I . As we will see below, any *other* paraxial ray originating from A will also give rise to a refracted ray in the second medium for which the ray path will intersect the axis at the *same* point I . This point (I) will then constitute the *image* of the object point A . One other such ray, for which the refracted ray path is easy to obtain, is AP , i.e., the ray incident at the pole P . Since this ray is normal to the tangent at P , it will be refracted without deviation, along PC , intersecting the refracted ray path from N at I and conforming to the above prediction.

We once again define the object distance (u) as the signed distance *from* the pole P to the object point A and, similarly, the image distance (v) as the signed distance from P to I . Following our sign convention elaborated earlier (sec. 10.9), one observes that, for the situation depicted in fig. 10-36, u is negative while v is positive. For given values of the refractive indices (n_1, n_2) of the two media separated by the refracting surface, the relation between the object distance, image distance, and the radius of curvature works out to

$$\frac{n_2}{v} - \frac{n_1}{u} = \frac{n_2 - n_1}{r}. \quad (10-36)$$

One may take note of the following features of the above relation.

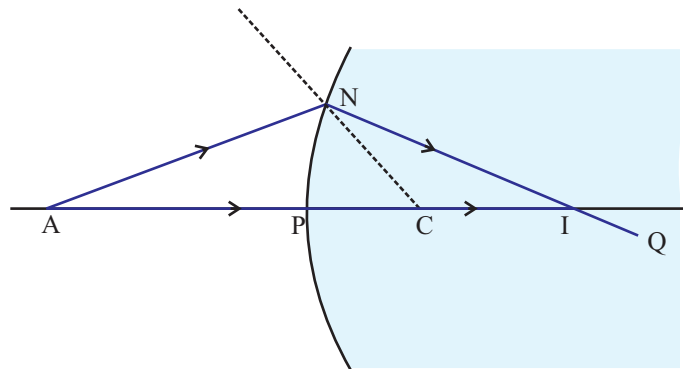


Figure 10-36: Image formation by refraction at a convex spherical surface; I is the (real) image point for an object point A situated on the axis PC.

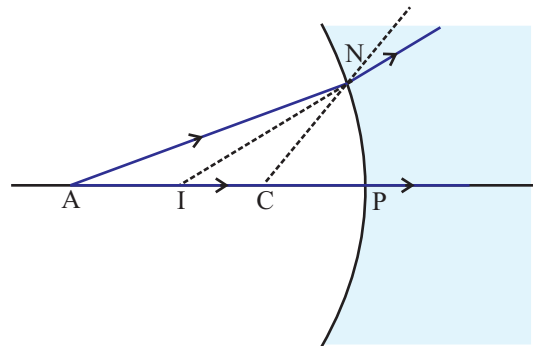


Figure 10-37: Image formation by refraction at a concave spherical surface; I is the (virtual) image point for an object point A situated on the axis PC.

1. The relation (10-36) between the distances PA, PI, and PC (note that all these are meant to be signed distances), which can be arrived at by considering the paraxial ray AN and the corresponding refracted ray path, does not contain any reference to the height of the point of incidence N above the axis or, equivalently, to its angle of incidence. The only assumption the derivation of the relation is based on (I include this derivation as an exercise below), is that the ray be paraxial. This means that any other paraxial ray would correspond to the *same* value of the distance PI as does the ray AN considered in the figure. In other words, *all* paraxial rays originating from A will give rise to refracted ray paths passing through I. This is why the point I has been termed the image point for object point A, and PI has been termed the image distance (v) corresponding to the object distance u . These ideas are, however, by now familiar to us.

2. While the relation (10-36) refers to a particular situation, namely the one shown in fig. 10-36, where the object point and its image point are on opposite sides of a convex refracting surface, it is of more general validity, subject only to the condition of the rays being paraxial. For instance, fig. 10-37 shows a concave refracting surface and a situation where the refracted ray path NQ corresponding to a paraxial incident ray AN diverges away from the axis, and intersects the latter only on being produced backwards, the point of intersection being I. Other paraxial rays originating from A are also similar in that the refracted ray paths intersect the axis on being produced backwards, and all those points of intersection are once again the same, namely I.

One then has a *virtual* image at I as opposed to a real image shown in fig. 10-36. Nevertheless, the relation between the object distance PA, image distance PI, and the radius of curvature PC is once again found to be the same as (10-36), where u , v , r are now *all negative*, in accordance to our sign convention. Likewise, in all other situations involving refraction of paraxial rays at spherical refracting surfaces, (10-36) happens to describe the relation between object distance, image distance and radius of curvature for given values of n_1 , n_2 , with u , v , and r carrying appropriate signs.

3. Since the (second) focal point can be interpreted as the image for an object located at an infinite distance from the refracting surface on the axis, it corresponds to a special case of formula (10-36) where the substitution $u \rightarrow -\infty$ should yield $v \rightarrow f$, or in other words,

$$\frac{n_2}{f} = \frac{n_2 - n_1}{r}. \quad (10-37)$$

This gives the second focal length (or, in brief, the focal length) f in terms of the radius of curvature and the two refractive indices n_1 , n_2 . A similar expression can be obtained for the first focal length (f') as well (work it out). Using (10-37), one

can write the relation between the object distance and image distance in the form

$$\frac{n_2}{v} - \frac{n_1}{u} = \frac{n_2}{f}. \quad (10-38)$$

4. Strictly speaking, the refractive indices n_1 , n_2 in (10-36) or (10-37) should also carry appropriate signs. Recall that we have assumed that all distances are to be measured from left to right. With this convention, if a ray proceeds from left to right, in a medium, then the refractive index is to have positive sign while, for a ray proceeding in the opposite direction, the refractive index is to be *negative*. This is necessary in order that formulas like (10-36) or (10-37) may hold regardless of the direction of the ray in this medium or that. As a simple application, let us try to see what happens if we apply formula (10-36) to the case of *reflection* at a spherical surface where the ray is bent, not into a second medium, but into the first medium itself, proceeding in the opposite direction. This suggests that one has to substitute $n_2 = -n_1$ in formula (10-36). Indeed, on making this substitution, (10-36) leads precisely to (10-32), as expected.

A word on notation. In the above considerations, the symbol r has been used in two different contexts - once, to denote angles of refraction, and once again to denote radii of curvature. One way to avoid confusion would be to use the upper case letter R to denote a radius of curvature. However, even when the lower case symbol (r) is used, one can avoid confusion by picking up the meaning of the symbol from the context.

Problem 10-11

Work out the derivation of (10-36) for the situation depicted in fig. 10-37.

Answer to Problem 10-11

HINT: In the triangles ANC and INC, one has, respectively, $\frac{AC}{AN} = \frac{\sin i}{\sin \theta}$, and $\frac{IC}{IN} = \frac{\sin r}{\sin \theta}$, where θ stands for the angle $\angle NCP$. Making use of the paraxial approximation and of the law of refraction, this gives $\frac{AC \cdot IP}{IC \cdot AP} = \frac{n_2}{n_1}$. In these relations, the lengths of the respective line segments are all positive

quantities. Referring to the figure and making use of the sign convention of sec. 10.9, one can write the above relation as $\frac{(-u+r)(-v)}{(-v+r)(-u)} = \frac{n_2}{n_1}$, from which the required relation follows.

Problem 10-12

A narrow bunch of parallel rays is incident on the surface of a transparent spherical object immersed in a liquid, and emerges from it after two refractions as shown in fig. 10-38. If a point image is formed on the front surface of the sphere, what is the index of refraction (n) of the material of the sphere relative to the surrounding medium?

Answer to Problem 10-12

HINT: The image formed by the front surface of the sphere (radius of curvature $r_1 = R$, the radius of the sphere; this corresponds to a convex refracting surface) acts as the object with reference to the refraction at the rear surface (compare a similar line of reasoning pursued in sec. 10.12.1; the rear surface in the present instance is a concave one). Since the object distance for the first refraction is $u = -\infty$, the image distance is, from eq. (10-36), $v = \frac{n}{n-1}R$, where n is the required refractive index of the material of the sphere with respect to the surrounding medium (reason this out). For the refraction at the second surface ($r' = -R$), the object distance is $u' = v - 2R$, while the image distance is $v' = -2R$. For this second refraction, the first and the second media have refractive indices n_2 and n_1 respectively, where $n = \frac{n_2}{n_1}$. Making use of eq. (10-36) once again, one obtains $n = \frac{2}{3}$. Note that the image distance for the first refraction is $v = -2R$, implying that the image is a virtual one. However, it acts as a real object for the second refraction.

Problem 10-13

A point object A is placed on the axis of a concave refracting surface (radius of curvature $r = -0.3$ m) separating a medium of refractive index 1.5 (the medium of incidence) from air (refractive index 1), while a plane mirror is placed at a distance 0.2 m behind the refracting surface (i.e., in air), the plane of the mirror being perpendicular to the axis of the refracting surface. At what distance in front of the refracting surface should A be located so that its final image will be at A itself?

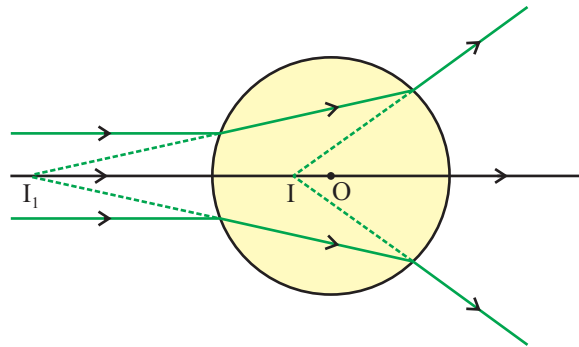


Figure 10-38: A parallel bunch of rays incident on a spherical object of refractive index n relative to the surrounding medium (we assume $n < 1$ for the sake of concreteness) and suffering two successive refractions; the image I_1 formed by the refraction at the front surface of the sphere acts as an object for refraction at the rear surface; the final image I is formed on the front surface if $n = \frac{2}{3}$ (refer to problem 10-12).

Answer to Problem 10-13

HINT: In order that the image may coincide with the object A, the rays, after refraction at the concave surface and the subsequent reflection at the plane mirror, have to retrace their path, which is possible only if the rays emerging in air after refraction at the concave surface are perpendicular to the mirror, i.e., are parallel to the axis of the refracting surface (draw your own figure, indicating a number of relevant ray paths). In other words, A should be located at the first focal point of the refracting surface (the distance between the refracting surface and the plane mirror is immaterial here). The required distance (u) from the refracting surface is thus obtained from $-\frac{1.5}{u} = \frac{1-1.5}{-0.3}$ (see formula (10-36), in which put $v \rightarrow -\infty$), i.e., $u = -0.9$ m.

10.11.2 Image formation for a short extended object

One can go on from here to address the question of image formation for an off-axis object point in much the same way that I related to you in section 10.10.3.2 in the context of reflection at a spherical surface. And then one can also look at the problem of image formation for a small extended object placed close to the axis. Without going into details, most of which would be redundant now, I illustrate in fig. 10-39 the formation of an inverted real image of a short extended object by refraction at a concave spherical surface, where the refractive indices of the media on the two sides of the surface satisfy

$$n_2 < n_1.$$

In this figure I show two ray paths originating from the point A on the object, where both the paths intersect at I, the image of the off-axis object point A. Denoting by u and v the horizontal distances of A and I from the pole P along the axis, you can make use of the figure to show that, once again, the relation (10-36) is satisfied. Interestingly, while the ray path ANI, with AN parallel to the axis, is to be paraxial, the other path ACMI *need not be so* for the derivation to hold. What is involved in the image formation is that a bunch of paraxial ray paths with their initial segments close to AN, all pass through I after refraction at the spherical surface. The path ACMI is simply a geometrical construct with the help of which one establishes the truth of this last statement.

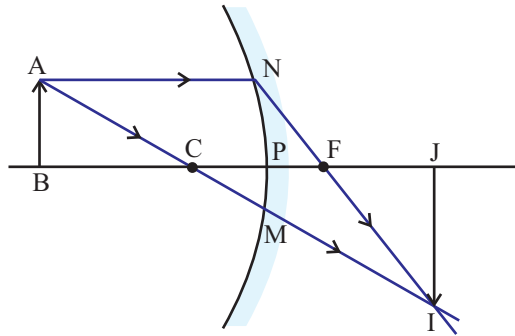


Figure 10-39: Image formation of a short extended object AB by refraction at a concave spherical surface; I is the (virtual) image point for an object point A situated on the extended object; the ray AN, incident at the point N on the surface, is refracted along NFI, where F is the second focal point; the ray AC passes undeviated along MI; I is the image point corresponding to A, and IJ is the image of AB; n_2 is assumed to be less than n_1 .

As in the case of reflection at a spherical surface, the principle of geometrical *similarity* between the object and the image holds here. Each small line element in the object is magnified by a constant ratio $\frac{n_1}{n_2} \frac{v}{u}$. Note that this reduces to the expression for magnification in the case of reflection at a spherical surface if one substitutes $n_2 = -n_1$, as indeed it should.

10.12 Spherical lens

Fig. 10-40 (A) shows a *lens*, a transparent object bounded by two refracting surfaces, at least one of which is curved. For the lens shown in the figure, both the refracting surfaces (S_1 and S_2) are curved and are convex with reference to light rays incident from outside, i.e., from the left and from the right respectively. Such a lens is termed a bi-convex (or, in short, *convex*) lens. Note from the figure that the central portion of the lens is thicker than the peripheral portions, the latter being tapering in comparison with the former. This is the general feature of a convex lens which may be of diverse shapes, depending on the nature of the two bounding surfaces. Thus, fig. 10-40 (B) shows a *plano-convex* lens, one of the two surfaces (S_1 in the figure) of which is convex while the other (S_2) is plane. On the other hand, the bounding surface S_1 in fig. 10-40(C) is convex while the other surface (S_2) is concave, making up a *concavo-convex* lens (sometimes referred to as a *meniscus* lens). By contrast, the central section of a *concave* lens is thinner compared to the upper and lower peripheral sections as in fig. 10-41 (A),(B),(C) showing a *double-concave*, a *plano-concave* and a *convexo-concave* lens respectively. Often, one refers to a double-concave lens as a concave lens in short.

In addition to spherical lenses considered in this section, *cylindrical lenses* are also in common use, where the lens surfaces are cylindrical ones. While the phenomenon of image formation by a cylindrical lens can be described in terms of the general principles of ray optics, a number of features of images formed by cylindrical lenses differ from those formed by spherical ones.

The line joining the centers of curvature of the two refracting surfaces of a lens is the common axis of both the surfaces and is termed the *axis* of the lens. Fig. 10-42 shows schematically the axis (C_1C_2) as also the poles (P_1 , P_2) of the refracting surfaces of (A) a convex and (B) a concave lens. The figure also shows, in a general way, the action of a convex and a concave lens on light rays incident on it from either side (one usually assumes that the rays are incident from the left, though there is nothing to prevent one

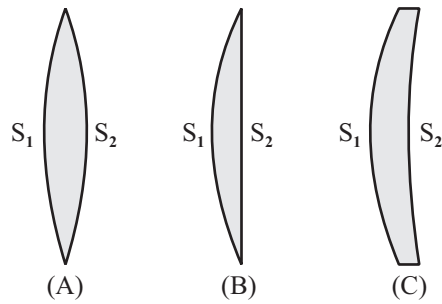


Figure 10-40: Convex lens of various types: (A) bi-convex, (B) plano-convex, (C) concavo-convex.

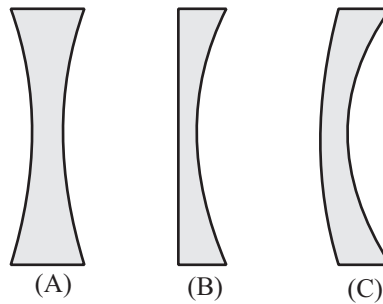


Figure 10-41: Concave lens of various types: (A) bi-concave, (B) plano-concave, (C) convexo-concave.

from considering rays incident from the right as well).

Observe that in fig. 10-42(A), a bunch of parallel rays incident on the lens is converted to a *converging* bunch of rays after being refracted at the two surfaces, while in fig. 10-42(b) the parallel bunch of rays is converted to a *diverging* one. In fig. 10-43(A) one finds a divergent bunch of rays being converted to a *less divergent* bunch on passing through a convex lens while in fig. 10-43(B) a convergent bunch of rays is converted into a parallel (more generally, to a *less convergent*) bunch on passing through a concave lens. In other words, a convex lens in general possesses a *converging* action on a bunch of rays while a concave lens possesses a *diverging* action. This is why convex and concave lenses are sometimes referred to as *converging* and *diverging* lenses respectively.

Fig. 10-44(A) will give you an idea why the convex lens possesses a converging action on a bunch of rays refracted through it. It represents the convex lens as a collection of

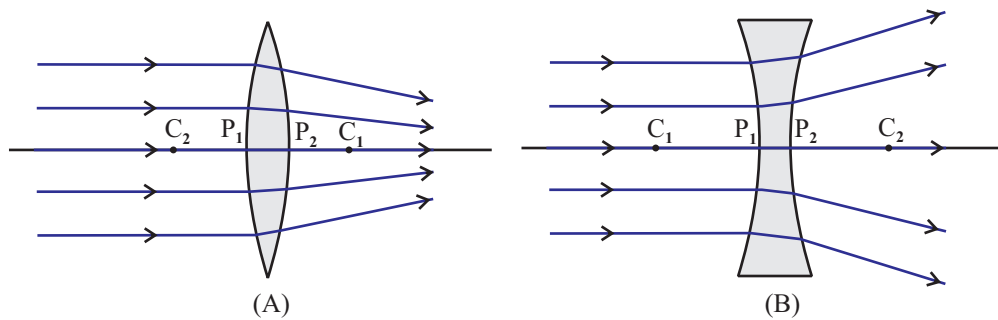


Figure 10-42: Poles (P_1 , P_2), centres of curvature (C_1 , C_2), and axis of (A) a convex lens, and (B) a concave lens; the action of either lens on a parallel bunch of rays is shown.

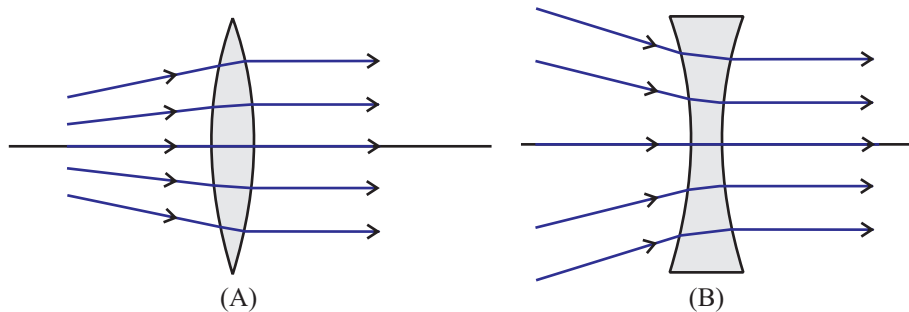


Figure 10-43: (A) Converging action of a convex lens, and (B) diverging action of a concave lens.

truncated prisms, with their bases all turned towards the axis. A truncated prism is one with a part of it cut off in such a way that it can still bend a ray at the two refracting surfaces. As mentioned in sec. 10.7 the ray *gets bent towards the base* of the prism by both the refracting surfaces. The amount of bending, measured by the deviation suffered by the ray, increases with the angle of the prism. In a similar fashion, fig. 10-44(B) shows a concave lens as a collection of truncated prisms, but now with the bases turned *away* from the axis.

Recall, for instance, the formula (10-19), where the deviation is found to increase in proportion to the angle of the prism. When a *thin lens* is imagined as a collection of truncated prisms, the angles of all these prisms are necessarily small, as a result of which (10-19) holds to a good degree of approximation, especially for paraxial rays. In the following we shall confine our attention to image formation by thin lenses.

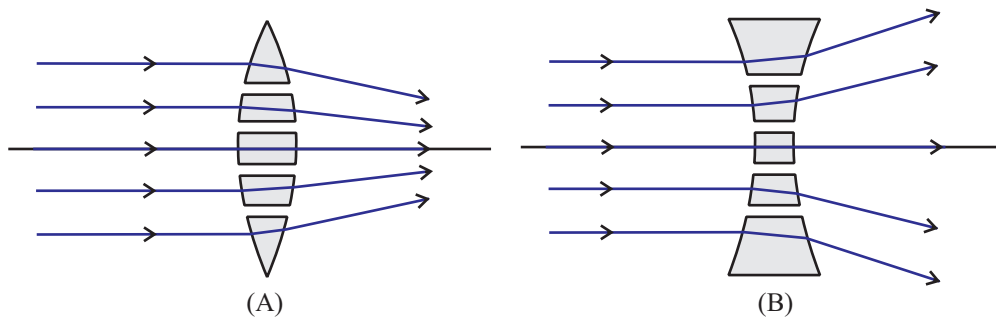


Figure 10-44: (A) A convex lens imagined as a collection of truncated prisms; the prisms have their bases turned toward the axis and the angles of the prisms increase from the axis outward; as a result the lens has a converging action on a bunch of rays incident on it; (B) a similar representation of a concave lens, explaining its diverging action.

Looking at fig. 10-44 you will observe that both for a convex and a concave lens, the angles of the truncated prisms increase as one moves from the axis towards the peripheral portions of the lens. This means that among a bunch of parallel rays, for instance, the ones incident near the periphery of the lens get bent to a greater extent compared to those incident near the axis, which explains the converging action of the convex lens as also the diverging action of a concave lens.

In the following we will confine ourselves to the consideration of *paraxial rays* alone because these are the rays for which the formulae relating to refraction at curved surfaces obtained in sec. 10.11 are valid.

Fig. 10-45(A), (B) define schematically the *first* and *second* principal foci of a convex lens. In 10-45(A), F_1 is a point on the axis such that a bunch of rays originating from F_1 is converted to a bunch of rays all parallel to the axis on passing through the lens. This point is termed the first principal focus (or, in short, the first focus) of the lens. Fig. 10-45(B) depicts the complementary situation where a bunch of rays parallel to the axis gets converted to a convergent bunch, converging to the point F_2 on the axis, the latter being termed the second principal focus (or, in short, the second focus; sometimes this is often referred to simply as the focal point or the focus of the lens). Fig. 10-46(A), (B) similarly define schematically the first and the second principal focus for a concave lens.

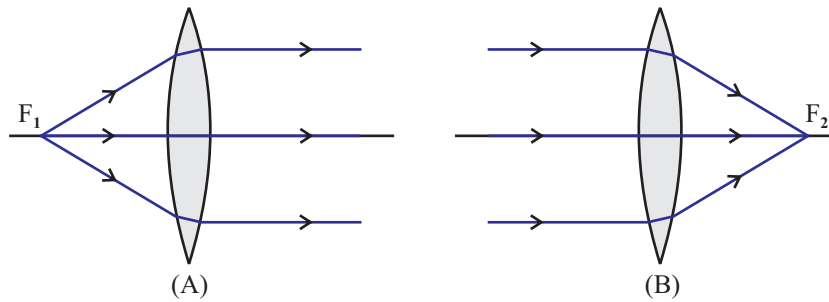


Figure 10-45: (A) First and (B) second focal points of a convex lens.

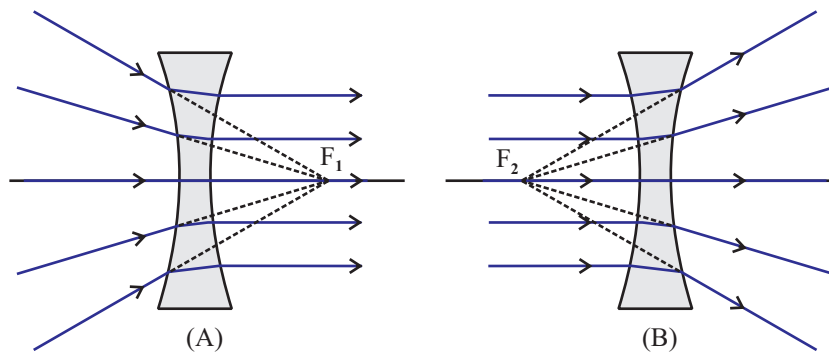


Figure 10-46: (A) First and (B) second focal points of a concave lens; note that, with reference to the ray paths, both the foci are virtual points.

A bunch of rays parallel to the axis can be imagined to have originated from or to meet at a point at an *infinite* distance located on the axis. In this sense, looking at 10-45(A), one can say that a set of rays originating from an object point at the first focus F_1 meet at an infinitely distant point on the axis after being refracted by the convex lens or, in other words, an object point at F_1 corresponds to an image point at an infinite distance.

By contrast, in fig. 10-46(A), F_1 corresponds to a *virtual* object since the rays that are converted to a parallel bunch by the concave lens do not actually originate from it. The image point here is, however, once again at an infinite distance on the axis. Thus the first focal point can be defined to be either the real or virtual object point for which the image point is located on the axis at an infinite distance.

Similarly, the parallel rays coming from an object point at infinity in fig. 10-45(B) con-

verge to F_2 which is thus the real image of the infinitely distant object while, by contrast, in 10-46(B), the rays from the object at infinity do not actually converge to F_2 but only appear to diverge from it which is thus a *virtual* image. One can thus define the second focal point as the real or virtual image point corresponding to an object point at an infinite distance.

We can now define a set of characteristic distances for a given convex or concave lens that will be found to be relevant in determining the way the lens forms real or virtual images for point objects, not necessarily located at an infinite distance. The equations giving the position of the image point corresponding to an arbitrarily specified position of the object point are known as object-image relations. The characteristic distances that determine the object-image relations for a given lens are its *radii of curvature* and *focal lengths*. These, however, are not independent quantities, but are related to one another, as we see below.

Consider the spherical surface of the lens on which the rays originating from the object point (or directed towards a virtual object point) are incident before being refracted by it and call it the *first* surface, while the other spherical surface through which the rays emerge after being refracted twice, will be referred to as the *second* surface of the lens. In the figures above, these are the left- and the right hand surfaces of the lens respectively.

Two of the characteristic distances mentioned above are then the *first* and *second* radii of curvature of the lens, which we denote as r_1 and r_2 . For instance, r_1 is the distance of the centre of curvature (C_1) of the first surface from the pole P_1 of that surface, while r_2 is similarly defined in terms of C_2 and P_2 (see fig. 10-42(A), (B)).

Notice that r_1 and r_2 are to be considered as quantities with appropriate *signs* in accordance with the sign convention we have already decided upon. Thus, for a bi-convex lens (fig. 10-42(A)) r_1 is *positive* since C_1 is located to the *right* of P_1 , while r_2 is *negative* since C_2 is located to the left of P_2 . On the other hand, for a concave lens (fig. 10-42(B)) r_1 is *negative* while r_2 is *positive*.

A convex lens for which the radii of curvature of the two surfaces are equal in magnitude (the signs of the two radii of curvature are opposite) is termed an *equi-convex* lens. An equi-concave lens is similarly defined. In other words, an equi-convex or an equi-concave lens is characterized by the relation $r_1 = -r_2$.

The two *focal lengths*, say, f_1 and f_2 , of a lens are defined in a manner analogous to the radii of curvature. Thus the first focal length (f_1) is the distance from P_1 to the first focus F_1 while the second focal length is similarly defined as the distance from P_2 to the second focus F_2 . As seen from fig. 10-45, 10-46, f_1 is negative for a convex lens and positive for a concave lens, while f_2 is positive for a convex lens and negative for a concave lens. It is customary to refer to a convex lens as a *positive* lens, and to a concave lens as a *negative* lens with reference to the sign of f_2 .

10.12.1 Image formation by a thin lens

We now address the question of image formation by *thin* lenses. A thin lens is one for which the poles for the first and the second refracting surfaces are close to each other so that it makes little difference if we assume the distance between the two to be zero. The two poles can then be assumed to coincide with the centre of the lens. In the following, we consider paraxial rays from point objects located on the axis refracted by the two lens surfaces in succession to form the corresponding images. Image formation for point objects located off the axis will be considered next.

Before we begin, I want to make it clear that, in any figure illustrating the bending of ray paths by a thin lenses, the latter will be shown to have a finite thickness, but the bending of the ray paths at the two faces of the lens will not be shown separately. Instead, the rays will be shown to be bent at a single plane, the central plane cutting through the lens (plane P in fig. 10-47). The poles P_1 , P_2 of the two curved faces of the lens, though shown as separate points in a figure, will be assumed to be coincident with the center C (to be distinguished from the symbol depicting the center of curvature of a spherical surface) of the lens. A ray passing through C will not be bent since a small portion of the lens near C acts effectively as a thin parallel-sided plate which produces

only a small lateral displacement without any change in direction, and for a thin lens this lateral displacement can be ignored.

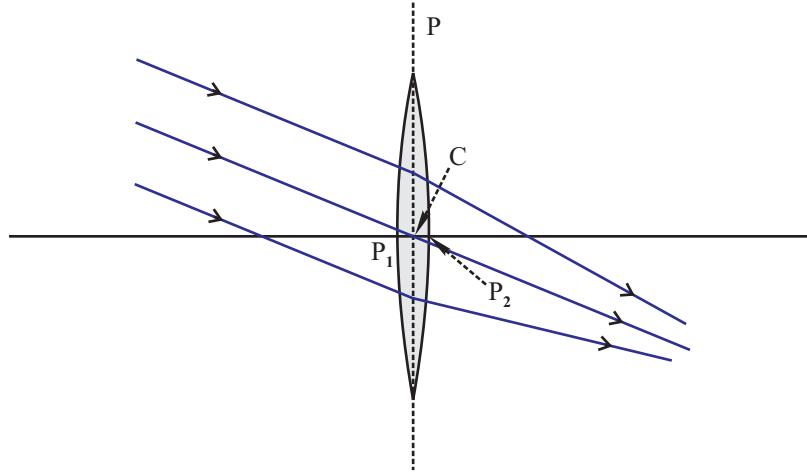


Figure 10-47: Explaining the way the ray paths through a thin lens are depicted; the poles of the two lens surfaces are close to each other; for the sake of convenience, the bending of a ray paths is not shown separately for the two lens surfaces; instead, the ray will be shown to be bent at a single plane, the central plane P, where the distances of the two poles from this surface may assumed to be vanishingly small; in between the two poles, the plane P cuts the optic axis at a point C (to be distinguished from the symbol depicting the center of curvature of a spherical surface) such that a ray through C may be assumed to pass undeviated.

Figure 10-48 depicts the way the image of a point object located on the axis of a thin lens is formed where a convex lens is seen to form a real image I of a point object A located on its axis. Referring to the figure, let the refractive indices of the media on the two sides of the lens be respectively n_1 and n_2 , the refractive index of the material of the lens being, say, n . A commonly encountered situation, where the lens is placed in air would then correspond to $n_1 = n_2 = 1$ to a good degree of approximation. Let the object-point A be located at a distance u , the *object distance*, from the centre (say, O, allowing for a slight change in notation) of the lens.

Imagine for the moment that the second surface is absent, the other side of the first surface being filled up with the lens material. In this imagined situation, let the image of the object point A, formed by the refraction of paraxial rays from this surface, be at the point I_1 at a distance, say, v_1 from O (which we assume to be coincident with

the poles P_1 , P_2 of the two lens surfaces). We would then have, in accordance with the object-image relation for a spherical refracting surface considered in sec. 10.11:

$$\frac{n}{v_1} - \frac{n_1}{u} = \frac{n - n_1}{r_1}. \quad (10-39)$$

The rays that would have converged to or appeared to diverge from this imagined image point I_1 are actually intercepted by the second surface of the lens where they suffer a second refraction. A convenient way of describing this second refraction is to look at I_1 as an effective object point with reference to the second refracting surface. The point I_1 can be seen to be a virtual or real object point for the second surface if it be a real or virtual image respectively for refraction at the first surface (check this out).

If the image formed by refraction at the second surface corresponding to the object point I_1 be located at I at a distance, say, v from the center (i.e., the common location of the poles P_1 and P_2) of the lens then I will be the image of the object point A *formed by the lens as a whole*. Invoking once again the object-image relation for refraction at a spherical surface, now for the second surface of the lens, we obtain:

$$\frac{n_2}{v} - \frac{n}{v_1} = \frac{n_2 - n}{r_2}, \quad (10-40)$$

(check out the reasoning leading to the relations (10-39) and (10-40)).

If we now add up the two equations ((10-39)) and ((10-40)), we obtain the following object-image relation for image formation by the lens as a whole:

$$\frac{n_2}{v} - \frac{n_1}{u} = \frac{n_2 - n}{r_2} - \frac{n_1 - n}{r_1}. \quad (10-41)$$

This is the object-image relation that we set out to derive, and the relations ((10-39)) and ((10-40)) were set up as intermediate steps for this derivation. Recall that, in this equation all the distances u (object distance), v (image distance), r_1 and r_2 (the two radii of curvature) are to be taken with appropriate signs. For instance, in fig. 10-48, where the locations of the object point (A), image point (I), and the intermediate image (I_1) are shown schematically in a particular instance of image formation by a convex lens, u and

v_1 are negative and v is positive, while r_1 and r_2 are positive and negative respectively.

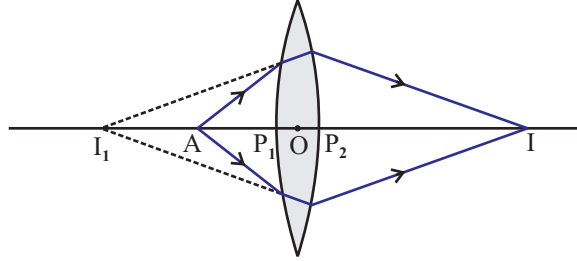


Figure 10-48: Image formation by a convex lens for a real point object A located on the axis; the image in this case is real; the points P_1 , P_2 , and O (the center of the lens) can be taken to be coincident for a thin lens; the intermediate image I_1 of A, formed by refraction at the first surface, acts as the object for the second surface, and gives rise to the final image I; the bending of the rays at each of the two surfaces is shown separately to indicate that the surfaces act in succession.

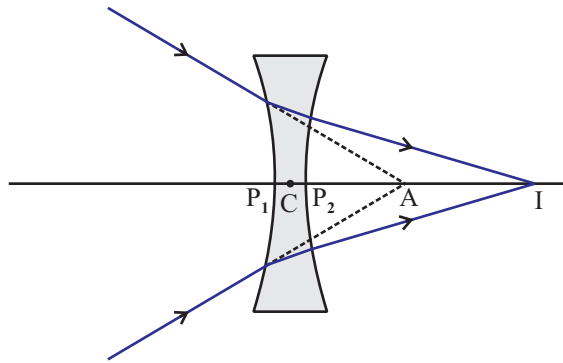


Figure 10-49: Image formation by a concave lens for a virtual point object (A) located on the axis; P_1 , P_2 , and C can be taken to be coincident for a thin lens; the intermediate image formed by the first surface is not shown; however, the bending of the rays at each of the two surfaces is shown separately.

While fig. 10-48 shows a convex lens for the sake of concreteness, there is nothing in principle in our derivation above that restricts the relation (10-41) to a convex lens alone. It is, in fact, applicable to image formation by a concave lens as well, provided only that the distances u , v , r_1 and r_2 are taken with appropriate signs. In other words, (10-41) is the general relation between object distance and image distance for image formation of a point object by a thin lens with paraxial rays, where we have assumed till now that the object is located on the axis of the lens.

A situation encountered quite frequently involves a lens made up of a material of refractive index n (say) placed in air where one can set, to a good degree of approximation, $n_1 = n_2 = 1$. In this commonly encountered situation (10-41) assumes the simpler form

$$\frac{1}{v} - \frac{1}{u} = -(n - 1)\left(\frac{1}{r_2} - \frac{1}{r_1}\right). \quad (10-42)$$

In the following, we will frequently refer to this simpler relation while considering image formation by a thin lens. The relation (10-41) will occasionally be referred to in the interest of generality.

The generality of the relations (10-41) or (10-42) subject to the conditions mentioned above allows us to use these even when the object under consideration is a *virtual* one as, say, in fig. 10-49, where a convergent bunch of rays is seen to be incident on a concave lens. These rays would converge to the point A in the absence of the lens and hence A is here a virtual point object, of which the image is formed at say, I. The latter is a real one for the situation shown in the figure since the emergent bunch of rays are seen to converge at, and thereafter diverge from, I. The fact that the object here is a virtual one is expressed in the sign of u which is *positive* (a real object would have corresponded to a negative u). The image distance v is also seen to be positive for the situation shown in the figure. While this sign could even be negative for a concave lens, a virtual object for a convex lens would correspond to positive signs for both u and v (check this out).

Let us now recall that the first focal length is the object distance corresponding to which the image distance is infinitely large and similarly, the second focal length is the image distance corresponding to an infinitely large object distance. Making use of (10-41) with $u = f_1$ and $v \rightarrow \infty$, i.e., $\frac{1}{v} \rightarrow 0$, (refer to figures 10-45(A), 10-46(A)), one obtains the following expression relating the first focal length to the two radii of curvature of the lens:

$$\frac{n_1}{f_1} = -\left(\frac{n_2 - n}{r_2} - \frac{n_1 - n}{r_1}\right), \quad (10-43)$$

where this relation holds both for a convex and a concave lens.

Similarly, with $\frac{1}{u} \rightarrow 0$, $v = f_2$ (figures 10-45(B), 10-46(B)) in (10-41), the second focal length is obtained in terms of the constants of the lens and the refractive indices of the three media as,

$$\frac{n_2}{f_2} = \left(\frac{n_2 - n}{r_2} - \frac{n_1 - n}{r_1} \right). \quad (10-44)$$

Specializing to the commonly occurring situation corresponding to $n_1 = n_2 = 1$ (approx.), the above two relations assume the simpler form

$$\frac{1}{f_2} = -\frac{1}{f_1} = -(n - 1) \left(\frac{1}{r_2} - \frac{1}{r_1} \right). \quad (10-45)$$

Finally, with $n_1 = n_2 = 1$, the object-image relation (10-42) can be written in the form

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f}, \quad (10-46)$$

where f stands for the second focal length (f_2) of the lens and is referred to simply as the *focal length*. The reciprocal of the focal length f is commonly referred to as the *power* of the lens.

1. For the formula (10-46) to hold, it is actually sufficient that the medium of incidence and the medium of emergence be the same, without either of these having the refractive index unity.
2. A great simplification occurs in the formulae of geometrical optics if one makes use of *reduced* distances rather than the distances themselves. For any distance (say, x) measured along the axis of an optical system, the reduced distance (x_r) is obtained by dividing with the refractive index (n) of the medium in which the distance is measured (thus, $x_r = \frac{x}{n}$).

In the case of image formation by a lens, if one makes use of the reduced distances, then the relation (10-46) holds even when the refractive indices of the object medium and the image medium are different from unity, provided u , v , and $f (= f_2)$ stand for the reduced object distance, the reduced image distance, and the

reduced second focal length. Denoting by f_1 and f_2 the two reduced focal lengths, one further finds that the relation $\frac{1}{f_2} = -\frac{1}{f_1}$ holds. For a medium of refractive index unity the reduced distance x_r is the same as the ordinary distance (x).

Problem 10-14

In the case of image formation by a thin lens of focal length f , if the object distance and image distance, *as measured from the first and second focal points*, be x and x' respectively, show that

$$xx' = -f^2. \quad (10-47)$$

Answer to Problem 10-14

HINT: In our derivations above the object distance u is the distance from the lens to the object while the distance from the lens to the first focal point is $-f$ (we assume that the medium of incidence and the medium of emergence are the same, in which case one has, $f_1 = -f_2 = -f$). Hence $x = u - (-f) = u + f$. In a similar manner, one has $x' = v - f$. Making use of these relations in eq. (10-46) we get $\frac{1}{x'+f} - \frac{1}{x-f} = \frac{1}{f}$, from which the required relation follows.

The formula (10-47) expressing the relation between the object distance and the image distance, measured from the two focal points, is referred to as Newton's formula.

10.12.2 Real and virtual image formation by a convex lens

Figure 10-50 depicts the formation of (A) a real image and (B) a virtual image by a convex lens. In fig. 10-50(A), the rays after emerging from the second surface of the lens converge to I, while in fig. 10-50(B), the ray paths, when produced backward, meet at I, i.e., in other words, the emergent rays appear to diverge from I. Evidently, for real image formation, the sign of the image distance v is positive, while in virtual image formation, the sign is negative.

For a real object point located on the axis, for which u is negative, the condition for virtual image formation by a convex lens is, $-u < f$, i.e., the object point is to be located closer to the lens than the first focal point (recall that, with $n_1 = n_2 = 1$, which we assume to hold for the sake of simplicity, $f_1 = -f_2$; more generally the relation $f_1 = -f_2$ holds when one refers to *reduced* focal lengths).

If, on the other hand, the object point is located further away from the lens than the first focal point, then the image will be a real one. More precisely, if $f < -u < 2f$, the image distance satisfies $v > 2f$, while for $2f < -u$, one finds $f < v < 2f$.

Finally, if the object be a virtual one, then the image formed by a convex lens will be real (check this out).

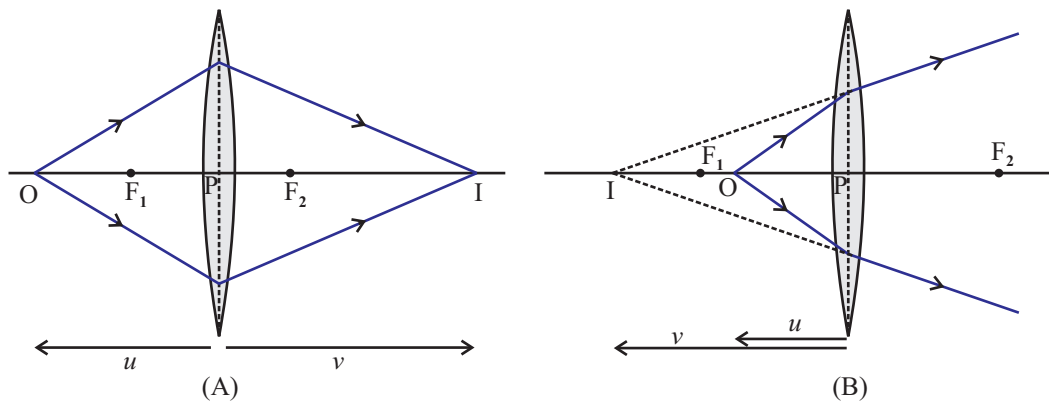


Figure 10-50: (A) Real image and (B) virtual image by a convex lens for a point object located on the axis.

10.12.3 Image formation for off-axis points

Fig. 10-51(A) shows a point object O located a short distance (y_O) above the axis of a convex lens. Two rays originating from O are shown in the figure. Of these, one ray is incident on the lens along a path (OA) parallel to the axis. On being refracted by the lens, the emergent ray passes through the second focal point (F_2) along the path AF_2 . Another ray, incident along the path AP at the optical center P of the lens (a slight change in notation once again, which I regret), proceeds undeviated and emerges along

PI.

The optical center of a thin lens is a point coincident with the two poles of the lens. A short segment of the lens on either side of the optical center can be thought of as a thin parallel-sided slab. A ray incident on the slab on one face comes out of the other face in a direction parallel to the incident ray, suffering only a small lateral displacement. Ignoring the latter, the ray paths for the incident and emergent rays can be taken to be along the same straight line.

The two emergent rays intersect at the point I on the other side of the lens, thereby forming the real image of the object point O. In this instance of real image formation by a convex lens of an off-axis point, the image point is seen to be located on the negative side of the axis where the object point is located on the positive side (refer to sec. 10.9; thus in the present instance, the distance (y_O) of the object point from the axis is positive, while the distance (y_I) of the image point is negative). In the figure, OM and IN are perpendiculars dropped from O and I respectively on the axis. The distance PM and PN measured from P to the points M and N respectively are referred to as the object distance (u) and the image distance (v).

As seen from the figure, the triangles OPM and IPN are similar to each other while, similarly, the triangles AF_2P and IF_2N are similar to each other. Making use of the geometrical properties of similar triangles one arrives at the following relation:

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f}, \quad (10-48)$$

which is the *same* relation as eq. (10-46) derived for an on-axis point through a slightly different approach.

Problem 10-15

Check out eq. (10-48).

Answer to Problem 10-15

HINT: $\frac{MP}{PN} = \frac{PF_2}{F_2N}$, both being equal to $\frac{OM}{IN}$, where the lengths of the respective segments are defined as positive quantities. Recalling that we are referring to fig. 10-51(A) as a particular instance, one has $MP = -u$, while the other segment lengths can be similarly related to u, v and f , from which the relation (10-48) follows. The same relation is seen to hold for figures 10-51(B) and (C) as well; see below.

In arriving at this relation, though we have considered only one pair of rays originating from O, in reality *all* paraxial rays originating from O and incident on the first surface of the lens intersect at I after emerging from the second surface. Indeed, this is why the point I is identified as the *image* of the point O.

Moreover, the relation (10-48) is seen to be independent of the height y_O of the object point above the surface of the axis, provided only that this height be sufficiently small. The latter is a necessary condition for the rays involved in the image formation to be paraxial ones. In other words, all points on any short segment of line OM perpendicular to the axis will be imaged at corresponding points on the segment IN, the latter being also perpendicular to the axis. Thus, IN may be termed the image of the extended object OM. What is more, one has from the triangles OPM and IPN,

$$m \equiv \frac{y_I}{y_O} = \frac{v}{u}, \quad (10-49)$$

where all the distances in the above equation carry their appropriate signs.

Thus the ratio of the transverse object distance (i.e., the distance of an off-axis object point from the axis) and the transverse image distance is the same for all object points located on the short segment OM. In other words, there is a *geometrical similarity* between a short extended object and its image formed by the lens. Though, in figure 10-51(A), OM is a line segment, all our conclusions hold for off-axis object points lying in any transverse plane, perpendicular to the axis. In other words, a small planar object placed perpendicular to the axis gives rise to a geometrically similar planar image in a

transverse plane.

Fig. 10-51(B) depicts the formation of a virtual image IN corresponding to a short linear object OM placed perpendicular to the axis of a thin convex lens, where the object is placed closer to the lens compared to the first focal point F_1 . Once again, it suffices to consider two rays originating from any point (say, O) on the extended linear object so as to construct its image. Analogous to the case of a real image, a small planar object placed perpendicular to the axis gives rise to a geometrically similar planar image perpendicular to the axis. The relations (10-48) and (10-49) are of general validity and hold in both the situations, provided all the distances are taken with appropriate signs.

Analogous statements apply for the image formation by a concave lens, one instance of which is shown in fig. 10-51(C). Once again, the relations (10-48) and (10-49) apply.

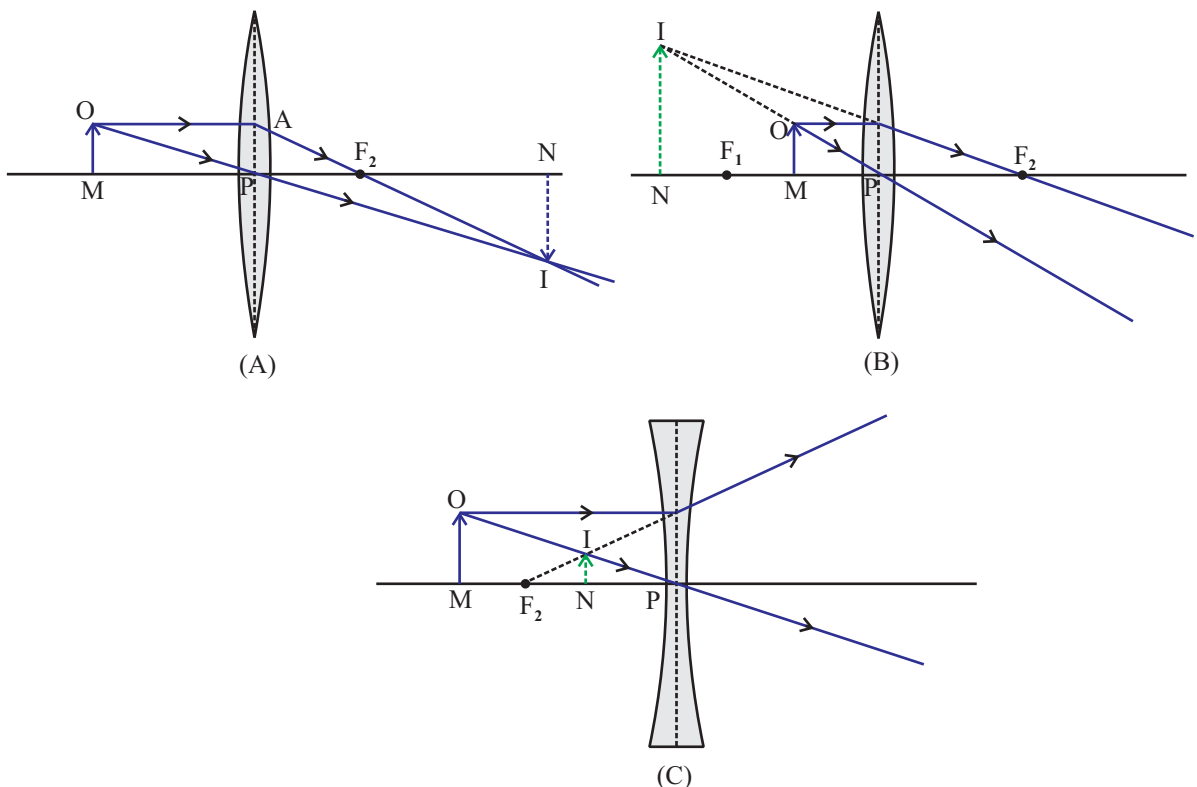


Figure 10-51: Image formation for an off-axis point and for a short linear extended object; (A) real image by convex lens; (B) virtual image by convex lens; (C) concave lens.

10.12.4 Longitudinal and transverse magnifications

Fig. 10-52 shows two point objects O_1 , O_2 , placed on the axis of a convex lens, at object distances, say, u and $u + \delta u$, where δu , the distance from O_1 to O_2 is positive (while u is negative) for the situation shown in the figure. Let the corresponding image points I_1 and I_2 be located at image distances v and $v + \delta v$ respectively. Assuming that the distances δu and δv are sufficiently small, one finds

$$\frac{\delta v}{\delta u} = \frac{v^2}{u^2}. \quad (10-50)$$

If one considers a short extended object O_1O_2 placed along the axis then its image will be a geometrically similar replica of the object, since distances along the axis are magnified in the ratio $\frac{v^2}{u^2}$, which is the same for all pairs of points in the object O_1O_2 , provided the latter is sufficiently short. This ratio, which depends on the location of the object on the axis, is termed the *longitudinal magnification* produced by the lens.

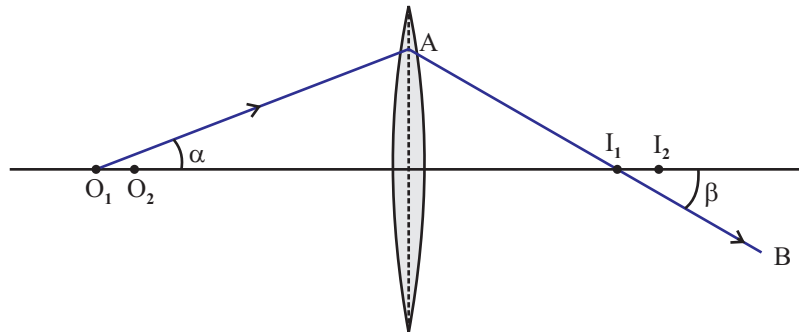


Figure 10-52: Illustrating longitudinal magnification produced by a lens; point objects placed at O_1 , O_2 on the axis, produce images at I_1 , I_2 ; the ratio $\frac{I_1I_2}{O_1O_2}$ is the longitudinal magnification, where the numerator and the denominator carry their own signs; the longitudinal magnification for a thin lens is always positive; the figure also illustrates the idea of angular magnification (see sec. 10.12.5), which is the ratio of the angles β and α , where the latter is the angle made by an incident ray with the axis, and the former is the corresponding angle for the emergent ray.

On the other hand, the ratio $m = \frac{v}{u}$ (see eq. (10-49)) is the *transverse magnification* (commonly referred to as the *magnification* in brief; the term linear magnification is used in order to distinguish it from angular magnification (see sec. 10.12.5)) produced by the

lens under consideration, which gives the factor by which a planar object perpendicular to the axis is magnified in giving rise to a geometrically similar planar image.

Note that, the transverse magnification m carries a sign of its own, since u and v do (in contrast, the longitudinal magnification is always positive). In particular, in the case of real image formation by a convex lens, the transverse magnification is *negative*. Thus, considering Cartesian axes MX, MY in the plane of the object, and parallel axes NX', NY' in the plane of the image (see fig. 10-53; the origins in the two sets of axes are the points M and N of fig. 10-51), the co-ordinates (x, y) of an object point and the corresponding co-ordinates (x', y') of its image point involve an *inversion* in the case of a real image.

One observes that, in general, the longitudinal and lateral magnifications are *not* the same. This means that the image of a small three-dimensional object will not, in general, be geometrically similar to the object.

Problem 10-16

At which point should a small three-dimensional object be placed in front of a convex lens so that it may give rise to a geometrically similar image?

Answer to Problem 10-16

HINT: The required condition is $v = -u = 2f$, since in this case the longitudinal and transverse magnifications are both of magnitude unity. The image under this condition is geometrically identical to the object except for an inversion about the origin in any plane perpendicular to the axis (in the sense indicated in caption to fig. 10-53).

10.12.5 Angular magnification

Fig. 10-52 also illustrates the concept of *angular magnification*. A ray O_1A originating from the object point O_1 , passes through the image point I_1 after refraction through the lens, along the path AI_1B . The angles made by the incident and emergent rays with the

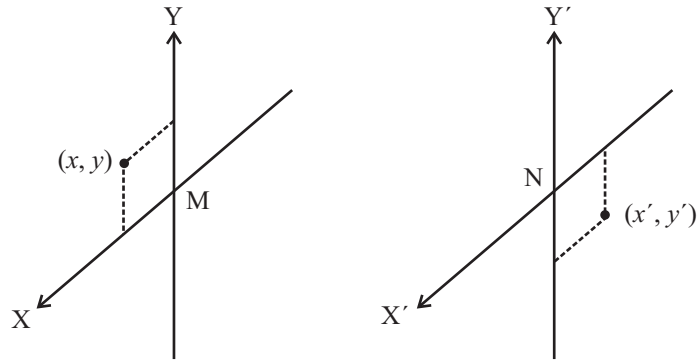


Figure 10-53: Illustrating planar inversion for a real image; N is the image of the on-axis point M of a lens; Cartesian axes MX, MY are chosen in a transverse plane containing M, while NX', NY' are a similar pair of axes in a transverse plane containing N; the image of the off-axis point (x, y) in the former plane is formed at (x', y') in the latter; for image formation by paraxial rays by a thin lens made of spherical surfaces, the transverse magnifications $\frac{x'}{x}$ and $\frac{y'}{y}$ are the same (independent of (x, y)); in the case of real image formation by a convex lens, an inversion in the transverse plane is involved, along with the magnification.

axis being, say α , and β , the angular magnification is defined as

$$\text{angular magnification} = \frac{\beta}{\alpha}. \quad (10-51)$$

Geometrical considerations relating to the triangles O_1AP and AI_1P give the following result

$$\frac{\beta}{\alpha} = \frac{u}{v}, \quad (10-52)$$

where the angles α and β have been assumed to be small so that the rays O_1A and AI_1B qualify as paraxial rays. Like the distances u and v , the angles α and β also carry their own signs. For the situation shown in figure 10-52, for instance, these two angles are, respectively, positive and negative (recall from section 10.9 the sign convention for angles).

Note from equations (10-49) and (10-52), that *the angular and the transverse magnifications are reciprocals of each other.*

10.12.6 Minimum distance between object and real image

Fig. 10-54 depicts a convex lens at position L_1 forming a real image of a short object (arrow pointing upward in the figure), located on the axis XX' , the image (arrow pointing downward) being at a distance of magnitude D from the object. If U , V , F be the magnitudes of the object distance, the image distance, and the focal length of the lens (i.e., $U = -u$, $V = v$, $F = f$ in terms of our earlier notation), then one has

$$\frac{1}{U} + \frac{1}{V} = \frac{1}{F}, \quad (10-53a)$$

where

$$U + V = D. \quad (10-53b)$$

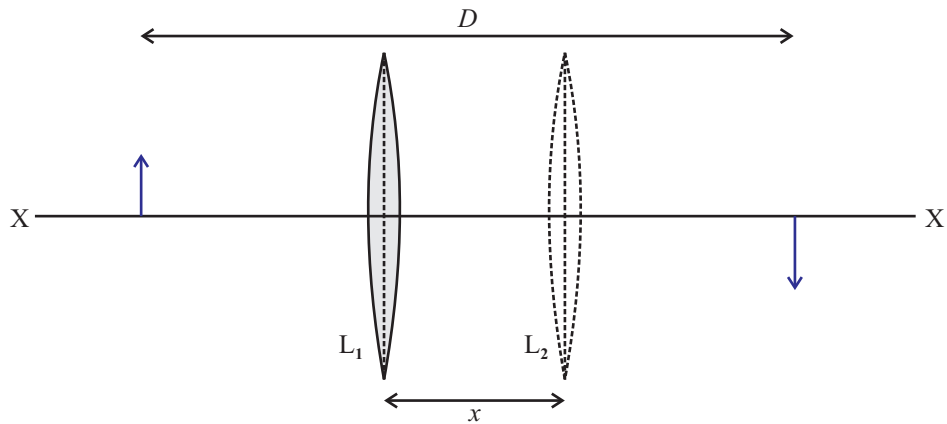


Figure 10-54: Illustrating the minimum distance between an object and its image in the case of formation of real image by a thin convex lens; for a given distance D between the object and the image, there are two positions of the lens for which a real image is formed, provided that D is larger than $4F$, where F stands for the magnitude of the focal length; for $D < 4F$, the lens cannot produce a real image of the object; for $D > 4F$, on the other hand, *two* real images are formed, corresponding to positions L_1 , L_2 of the lens, for the given distance between the object and its image.

Eq. (10-53a) is often invoked in working out problems relating to the formation of a real image by a convex lens (where, note that U , V , F are all defined to be positive quantities).

For a lens of a given focal length, one can eliminate V in favour of D in equations (10-53a), (10-53b) to set up a quadratic equation in U . Demanding that this equation should possess real roots for U , one obtains the condition

$$D \geq 4F. \quad (10-54a)$$

In other words, for real image formation by a convex lens, the *minimum possible separation between the object and the image* is four times the focal length of the lens. For given positions of the object and the image, with the value of their separation D larger than $4F$, there are *two* positions of the lens for which a real image can be formed, corresponding to the two real (and positive) roots of U of the quadratic equation referred to above. While one of these is at L_1 in fig. 10-54, the other position (L_2) is depicted with dotted lines. The separation between the two positions is found to be

$$x = \frac{D^2}{4F}. \quad (10-54b)$$

Problem 10-17

Set up the quadratic equation referred to above, and establish equations (10-54a), (10-54b). Show that $m_1 m_2 = 1$, where m_1 and m_2 are the transverse magnifications in the two positions of the lens.

Answer to Problem 10-17

HINT: Substituting $V = D - U$ in formula (10-53a), one obtains the required quadratic equation as $U^2 - DU + FD = 0$. The condition for the existence of two real roots for U is seen to be $D > 4F$, the two roots being $U_{1,2} = \frac{1}{2}(D \mp \sqrt{D^2 - 4FD})$, corresponding to positions $L_{1,2}$ in fig. 10-54. The corresponding values of V are given by $V_1 = U_2$, $V_2 = U_1$ (reason out why; make use of the fact that U and V are mutually conjugate variables, i.e., object and image positions are interchangeable; alternatively, one can make use of the quadratic equation in V obtained by eliminating U). The two magnifications being $m_1 = -\frac{V_1}{U_1}$, $m_2 = -\frac{V_2}{U_2}$ (note the minus signs), one obtains $m_1 m_2 = 1$.

Problem 10-18

In the case of real image formation by a thin convex lens, the magnitude of the distance from the object to the image is $D = 0.60$ m, while the magnification is $m = -2$. Obtain the values of the object distance u , the image distance v , and the focal length f of the lens.

Answer to Problem 10-18

HINT: The relevant relations here are $D = -u + v$ (reason this out), $m = \frac{v}{u}$ and the formula (10-46) relating u , v , and f , where all the quantities carry their own signs. Thus, $D = (m - 1)u$, giving $u = -0.20$ m, $v = 0.40$ m, and, finally, $f = 0.133$ m (approx).

10.13 Combination of thin lenses

Fig. 10-55 depicts a set-up involving *two* convex lenses L_1 and L_2 with an object AB placed in front of L_1 . The combination of the two lenses forms an image $A''B''$, where the formation of this image can be interpreted as the result of the formation of two images in succession. Rays originating from any point, say, A on the object, on being refracted by L_1 , converge at A' , thereby forming an intermediate image of A. Considering other points on the object, one obtains an intermediate image $A'B'$ formed by L_1 . This intermediate image $A'B'$ acts as an object for the lens L_2 . For instance, the rays originating from A that converged at A' , diverge thereafter and get refracted by L_2 , forming the final image A'' of A.

Thus, A' is the image of A formed by L_1 , and A'' is the image of A' formed by L_2 , while A'' is the image of A formed by the *lens combination* made up of L_1 and L_2 . Considering other points on the object AB, one obtains the image $A''B''$ formed by the lens combination.

While fig. 10-55 shows a combination of two convex lenses, one can have a combination of a convex and a concave lens or a combination involving two concave lenses. For instance, fig. 10-56 shows the formation of a virtual image by a combination of two concave lenses, where the intermediate image is also a virtual one. Combinations involving more than two lenses are also possible.

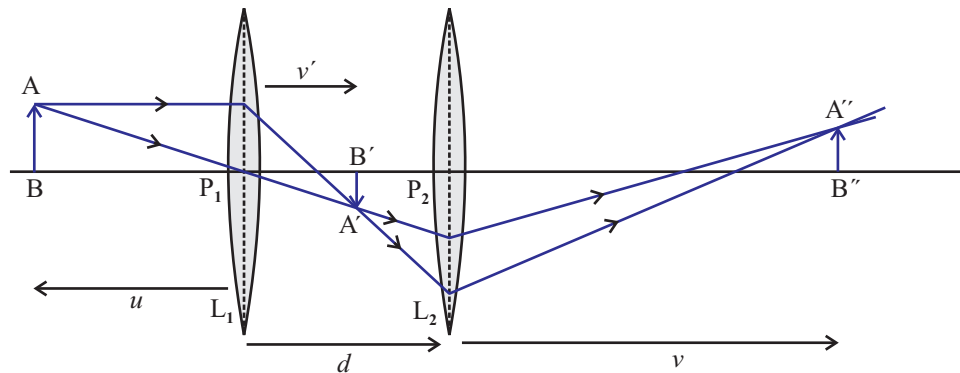


Figure 10-55: Image formation by combination of two convex lenses, L_1 and L_2 ; AB is a short object, placed perpendicularly to the common axis (P_1P_2) of the lenses; $A'B'$ is the image of this object formed by refraction through L_1 ; $A'B'$ acts as an object for the lens L_2 , and the final image is formed at $A''B''$.

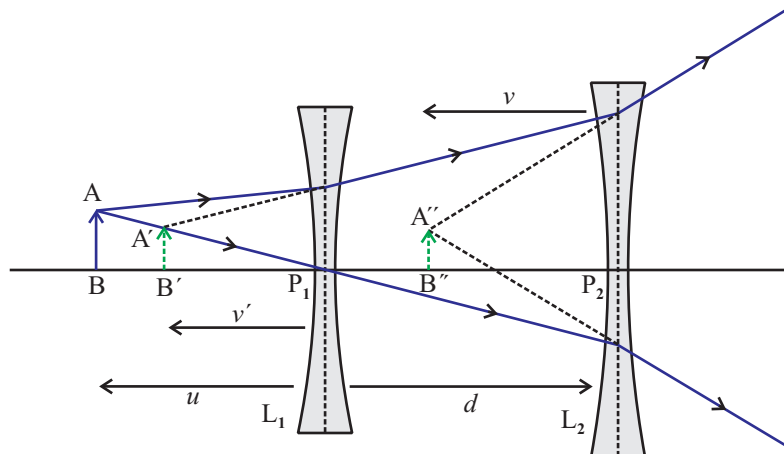


Figure 10-56: Image formation by a pair of concave lenses L_1 and L_2 ; the combination forms the image $A''B''$ of the short object AB placed in front of L_1 ; the intermediate image $A'B'$ formed by L_1 acts as an object for L_2 , giving rise to the final image $A''B''$; for two concave lenses in combination, both the intermediate and the final images are virtual; however, the virtual image $A'B'$ formed by L_1 acts as a *real* object for L_2 .

Interpreting the image formation by a lens combination in the above manner as a result of a succession of images formed by the individual lenses, where the image formed by one lens acts as the object for the next lens, one can work out the position of the final image in terms of the position of the object and the positions of the lenses in the lens combination.

For instance, considering a combination of two lenses of focal lengths f_1 and f_2 (note that the suffixes ‘1’ and ‘2’ refer here to the two lenses making up the combination, rather than to the first and second focal lengths of a single lens), let u denote object distance measured from the centre of the lens L_1 (the lenses being assumed to be thin ones, this is the point where the poles of the two lens surfaces may be assumed to coincide). Denoting by v' , d and v the distance of the intermediate image from the center of L_1 , the distance of L_2 measured from L_1 , and the distance of the final image from the center of L_2 respectively (see figures 10-55, 10-56), one has the following two equations, one for the formation of the intermediate image by L_1 , and the other for the formation of the final image by L_2 , with the intermediate image acting as the object:

$$\frac{1}{v'} - \frac{1}{u} = \frac{1}{f_1}, \quad (10-55a)$$

$$\frac{1}{v} - \frac{1}{v' - d} = \frac{1}{f_2}. \quad (10-55b)$$

In these equations all the distances carry their respective signs (see problem 10-19 for a concrete example). For instance, in fig. 10-55, for which f_1 and f_2 are both positive, u is negative while v' , d , and v are all positive. In fig. 10-56, on the other hand, for which f_1 and f_2 are both negative, u , v' and v are all negative, while d is positive.

10.13.1 Equivalent lens

In this context, the idea of an *equivalent lens* is sometimes useful. Consider a system of two lenses as, say, in fig. 10-55 or fig. 10-56. Imagine the pair of lenses L_1 and L_2 to be replaced with a *single* lens L of an appropriate focal length, say, f , and placed at an appropriate location such that the transverse magnification produced by L for a short object placed at any arbitrarily chosen point on the axis is the same as the magnification produced by the combination of the two lenses L_1 and L_2 . This means, in other words, that the single lens L is *equivalent* to the system made up of the lenses L_1 and L_2 from the point of view of magnification produced, for each and every position of the object. However, the *position* of the image formed by L will, in general, differ from that of the

image produced by the lens combination.

Thus, the equivalence between the single lens L and the combination of L_1 and L_2 is a limited one, being only in the sense of the magnification produced. In order that this limited equivalence be realized, the lens L has to have an appropriate focal length and is to be placed at an appropriate position with reference to the lenses L_1 and L_2 .

In other words, a combination of two thin lenses separated by a distance cannot, in any sense, be replaced with a single lens that can be considered as equivalent to the given combination in *all* respects. However, at times, a lens combination is replaced with a single lens for convenience of representation. The mathematical formulas written out for this single lens remain valid provided one interprets them appropriately.

For instance, when one speaks of the first focal plane of the combination and, along with it, of the first focal length (say, f_1), one has to imagine it as a plane located at a distance f_1 , *not* from either of the two lenses making up the combination, but from a certain plane referred to as the *first principal plane* of the combination. In the figures, however, it is commonly shown as a plane at a distance f_1 from a hypothetical equivalent lens for the sake of simplicity. A similar care has to be exercised in respect of the second focal plane and the second focal length f_2 , the latter being the distance from another particular plane, namely, the *second principal plane* of the combination though, once again, it is often depicted as the distance from a hypothetical equivalent lens. These qualifications are to be kept in mind while interpreting the mathematical formulas involving lens combinations.

A lens combination can be referred to as an effectively converging or a diverging one depending on how it bends the ray paths passing through it. For a converging combination the signs of f_1 and f_2 will be, respectively, negative and positive as for a single converging lens and if the lens combination be placed in air one will have $f_1 = -f_2$. Similarly, for a diverging lens combination, f_1 will be positive while f_2 will be negative and one will have $f_1 = -f_2$, assuming that the combination is placed in air. Similarly, other formulas involving f_1 , f_2 will remain valid if interpreted appropriately. The second

focal length will be commonly referred to as the focal length of the combination.

Incidentally, the suffixes '1' and '2' in the symbols f_1 and f_2 in this section refer to the first and second focal lengths of the combination, differing thereby from the notation in sec. 10.13 where f_1 and f_2 referred to the two lenses making up the combination.

Problem 10-19

Two thin lenses L_1 , L_2 are placed co-axially, with L_2 at a distance $d = 0.18\text{m}$ from L_1 (see fig. 10-57); the focal lengths of L_1 (concave) is $f_1 = -0.20\text{m}$, while that of L_2 (convex) is $f_2 = 0.1\text{m}$; if an object is placed at a distance $u_1 = -0.40\text{m}$ from L_1 , where will the image be formed by the combination, and what will be its magnification? If a single lens is to be used to form an image of the object at the same position as that produced by the combination, and with the same magnification, at what distance should it be placed from L_1 ? What should be the focal length of this lens?

Answer to Problem 10-19

The distance v_1 of the first image (i.e., the intermediate image formed by L_1) from L_1 is obtained from $\frac{1}{v_1} - \frac{1}{u_1} = \frac{1}{f_1}$, i.e., $v_1 = \frac{u_1 f_1}{u_1 + f_1}$, and the magnification is given by $m_1 = \frac{v_1}{u_1} = \frac{f_1}{u_1 + f_1}$. Substituting given values, one obtains $v_1 = -0.133$ (numerical results are approximate; all distances in meter), and $m_1 = 0.333$. The distance from this intermediate image from L_2 , for which it acts as the object, is $u_2 = v_1 - d = -0.313$ (refer to fig. 10-57), where all distances carry their appropriate signs. The distance of the final image from L_2 is then $v_2 = \frac{u_2 f_2}{u_2 + f_2}$ (refer to eq. (10-46), where the notation is different) which, on substituting relevant values, works out to $v_2 = 0.147$, from which the magnification produced by L_2 is obtained as $m_2 = \frac{v_2}{u_2} = -0.47$. The resultant magnification, produced by the combination of L_1 and L_2 is then $m = m_1 m_2 = -0.157$. The distance of the final image C from L_1 is $v_2 + d$ (see, once again, fig. 10-57).

Let the distance of the lens L (which, acting singly, produces the image of A at the same point C and with the same magnification m as does the combination of L_1 and L_2) be x . The distance of the object A from L is then $u = u_1 - x$, while the distance of the image C from L is $v = v_2 + d - x$. The magnification produced by L is thus $\frac{v_2 + d - x}{u_1 - x}$, which is to be equated to $m = -0.157$, giving $x = 0.228$. making use of this value of x , one obtains $u = -0.628$, $v = 0.099$. With this value of the object distance and the image distance, the focal length of L is obtained as $f = \frac{uv}{u-v} = 0.0855$.

In other words, the lens L , which is equivalent to the combination of L_1 and L_2 for the given position of the object (i.e., A) is a convex one which is to be placed beyond L_2 , as in fig. 10-57. The equivalence, however, obtains only for this particular position of the object.

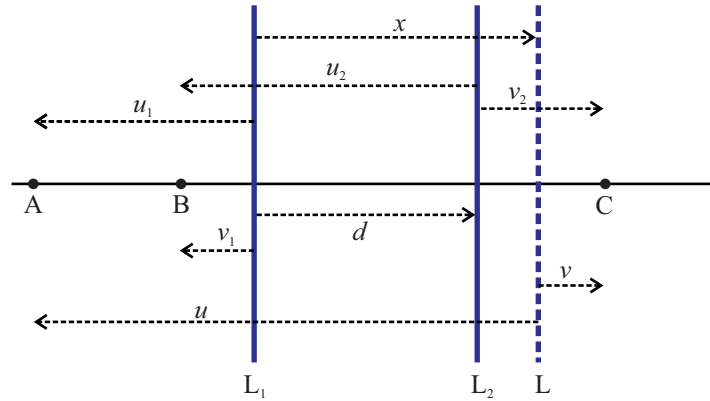


Figure 10-57: Image formation by a coaxial combination of a concave lens L_1 and a convex lens L_2 (only the planes indicating lens positions are shown); for an object A placed at a distance u_1 from L_1 , the combination produces an image C at a distance v_2 from L_2 (other distances are also marked; B is the intermediate image formed by L_1), with a magnification m ; a single lens L with an appropriate focal length f , placed at an appropriate distance x from L_1 , forms an image of A at the same position C and with the same magnification m , as worked out in problem 10-19; the position and magnification of the image produced by L will, however, differ from that produced by the combination, for any other position of the object A .

10.13.2 Thin lenses in contact

As a special case of combination of thin lenses, consider two lenses L_1 and L_2 placed in contact with each other. This means, in terms of fig. 10-55 or fig. 10-56, say, that $d = 0$. One can assume the centers of the two lenses to be coincident at, say, C . If now, a single lens L be placed at C in lieu of the pair of lenses, it will be equivalent to the latter if its focal length f is chosen to satisfy

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2}, \quad (10-56)$$

i.e., if its power is the sum of the powers of the two lenses L_1 and L_2 (here we revert back to the notation of sec. 10.13 where f_1, f_2 denote the focal lengths of the two lenses in contact, in contrast to the notation in sec. 10.13.1). This is easily seen by adding up

equations (10-55a) and (10-55b) which, with $d = 0$, results in

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f_1} + \frac{1}{f_2}, \quad (10-57)$$

where u and v are measured from the common center C, and which shows that the lens L with focal length f given by eq. (10-56) placed at C, will produce an image at the same position as that produced by the combination, for any *arbitrarily* chosen position of the object. What is more, the magnification produced by L will also be the same as that produced by the combination (check this out). In other words, the equivalence between L and the combination of L_1 and L_2 is not a limited one in this special case of L_1 and L_2 being in contact, but holds in respect of both the position of the image and the magnification produced.

10.14 Aberrations in image formation

Optical *systems* involving lenses, aperture stops, and (as necessary) mirrors, are used for purposes of image formation where angular or linear magnification is to be achieved and a faithful optical recording of the object is to be obtained. Fig. 10-58 show a lens L with an aperture stop S, along with a point object O, not necessarily located on the axis of the lens, the point being contained in a plane P_1 perpendicular to the axis (the 'object plane'). An aperture stop is commonly a circular aperture in an opaque screen placed in front of or at the back of a lens so as to limit the angular divergence of the bunch of rays that can pass through the optical system. Such stops are used to enhance the quality of image formation by the system under consideration.

The figure shows a conical bunch of rays allowed by the aperture stop, where all the rays belonging to the bunch are not necessarily paraxial, since restriction to paraxial rays alone would diminish the brightness of the image. As a result, after passing through the lens (or, more generally, through an optical system) the rays do not all converge to a single point I (the point where the ideal image is expected to be formed; in the present instance, the image formed by paraxial rays) in a conjugate 'image plane' perpendicular to the axis. The bunch of rays originating from O may be imagined to be made up

of small subgroups of rays, where each subgroup, after passing through the optical system, converges to a point that differs from the ideal image point I . In other words, there results a number of narrow pencils of emergent rays where these converge at various different points, say, I_1, I_2, \dots , and diverge thereafter. It is this deviation from ideal image formation for an object point that is shown schematically in fig. 10-58.

Considering an observation or recording plane P_2 , say, a plane containing the ideal image point I (which we take to be the image formed by a narrow cone passing through the center of the lens), one then gets, in general, a blurred set of illuminated points instead of a single sharp image point. Moreover, for an extended object, the geometrical similarity between the object and the image is lost to some extent.

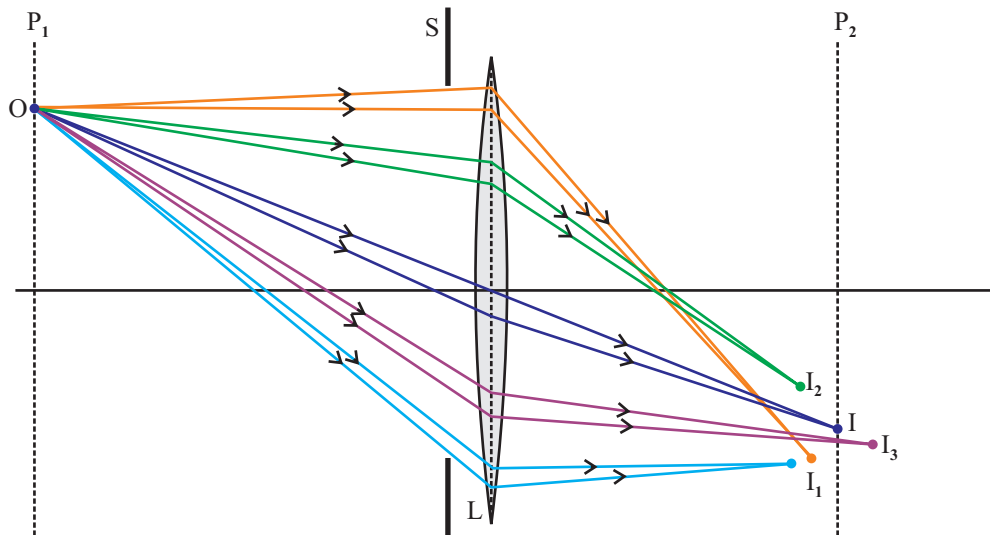


Figure 10-58: Explaining the idea of monochromatic aberration; a single lens with an aperture stop is shown; a cone of rays passing through the optical system can be imagined to be made up of a large number of small subgroups, each corresponding to a much narrower cone, where each of the cones produces its own image point, causing deviation from ideal image formation (schematic).

10.14.1 Monochromatic and chromatic aberrations

Such deviation from ideal image formation by an optical system is referred to as *aberration*. Optical systems usually produce aberrations of various different *types*, each type

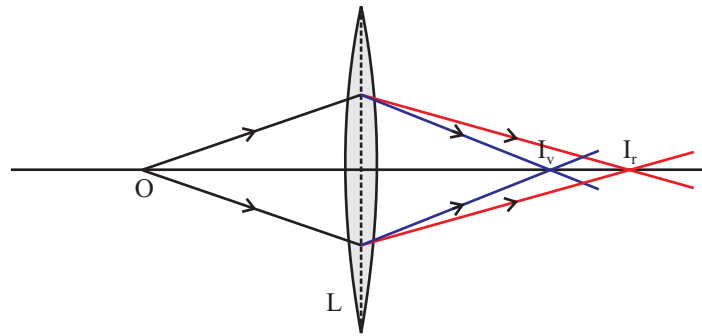


Figure 10-59: Chromatic aberration (schematic); even when a single narrow cone of incident rays is considered, a single point image is not formed since optical radiation of different colors gives rise to different images I_v , I_r .

being characterized by its own distinguishing feature. To start with, one identifies two broad types, namely *monochromatic* and *chromatic* aberrations. The mechanism I have described in the above paragraphs is the one essentially responsible for the monochromatic aberrations since this mechanism causes a deviation from ideal image formation even when the radiation sent out by the object point is a monochromatic one, i.e., is characterized by a single frequency.

Supposing, on the other hand, that the light sent out from the object point O is made up of components with more than one frequencies, there occurs a deviation from ideal image formation *even for a very narrow pencil of rays* passing through the optical system. Such a deviation, referred to as *chromatic aberration*, arises by virtue of the fact that the refractive index of light varies with the frequency, as a result of which light of various different colors follow different paths on being refracted by a lens. Fig. 10-59 shows a narrow pencil of rays from an object point O, but now there are two colors involved, say red (r) and violet (v). Since violet light is bent more than red light in refraction, one obtains two different image points (I_v) and I_r , and once again there occurs a deviation from a single ideal image point. The image recorded on any observation plane perpendicular to the axis will, in general consist of a colored patch instead of a sharp point.

10.14.2 Types of monochromatic aberration

Monochromatic aberrations, in turn, can be of several types, that have been classified as *spherical aberration*, *coma*, *astigmatism*, *distortion*, and *curvature*. Of these, the first three, namely, spherical aberration, coma, and astigmatism, belong to the category of *point-imaging aberrations*, where a point object does not give rise to a point image owing to non-paraxial rays being involved in the image formation, and the image of the point object appears blurred and spread-out. The other two, namely, distortion and curvature are aberrations involving the *image shape* for an extended object. In these cases, a point object gives rise to a point image, but the position of the latter differs from that of the ideal image produced by paraxial rays. As a result, the geometrical similarity between an extended planar object and its image is lost.

As an illustration of a monochromatic aberration, fig. 10-60 illustrates how a single lens may cause *spherical aberration* to occur. Paraxial rays originating from the object point O on the axis of the lens converge to the point I, while *peripheral* rays, incident at points near the rim of the lens, emerge from it to converge at a different point I'. The distance between the two images I and I' may be taken as a measure of the departure, caused by spherical aberration, from ideal imaging (more specifically, this distance is referred to as the *longitudinal* spherical aberration produced by the lens). If the defective image formed by the lens is recorded on the plane P shown in the figure, a blurred circle will be observed. The diameter of the circle, known as the circle of least confusion, can be taken as another measure (the *transverse* spherical aberration) of the extent to which spherical aberration is introduced by the lens. The circular patch recorded on any other plane of observation will have a larger radius.

10.14.3 Aberrations: an overview

While chromatic aberration arises due to the dependence of the refractive index of a material on the wavelength of light and the consequent dependence of a ray path through an optical system on the color of light, the monochromatic aberrations arise due to the deviation of the ray paths from paraxiality.

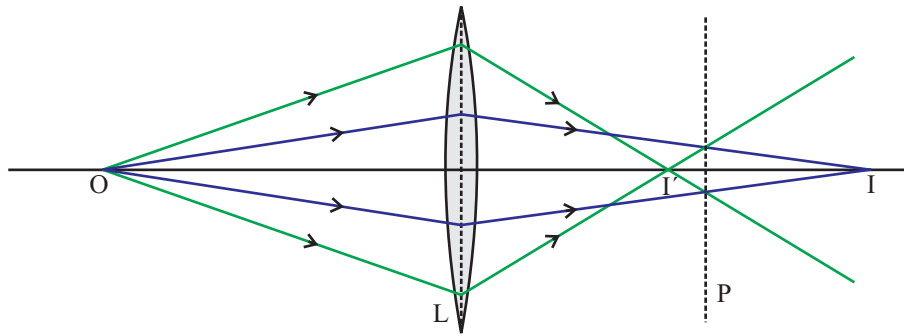


Figure 10-60: Illustrating spherical aberration produced by a convex lens (schematic); I and I' are the paraxial and peripheral images of the point object O located on the axis; the image recorded in the plane P is a circular patch rather than a sharp point.

I have already explained in a general way the terms ‘paraxiality’ and ‘paraxial rays’ as, for instance, in the case of a spherical reflecting surface by means of conditions (10-20). In general, for an optical system made up of a set of co-axial spherical surface, the condition of paraxiality requires that squares and higher powers of angles made by the ray paths with the axis are to be negligible and, similarly, the squares and higher powers of distances from the axis are also to be negligibly small. These conditions specify the limits of what is known as *Gaussian optics* where, more generally, one can consider non-spherical but axially symmetric surfaces. The monochromatic aberrations can thus be described as defects in image formation due to deviations from the limits of Gaussian optics.

Aberrations can also be explained in terms of the wave theory of light. According to the wave theory, a point image is formed when a spherical wave front (the concept of a wave front has been explained in chapters 9 and 14) converges to a point. Monochromatic aberrations such as spherical aberration and coma can then be explained as deviations from the spherical shape of the wave fronts emerging from the optical system under consideration. Such deviations arise due to factors relating to the incident wave front as also to the optical system itself. Generally speaking, however, all the monochromatic aberrations in axially symmetric optical systems can be looked upon as defects caused by the transgression of the limits of Gaussian optics.

Chromatic aberrations, on the other hand, are caused by the phenomenon of dispersion.

10.14.4 Correcting an optical system for aberrations

An optical instrument like a telescope or a microscope is made of a number of lens combinations, where a lens combination consists of more than one lenses in contact or, more generally, separated by a distance. In designing the instrument, one has several parameters at one's disposal that can be adjusted for optimum performance. For instance, the types of glass of which the lenses are made, the radii of curvature of the surfaces of the lenses, and the distances between the lenses in a lens combination, can be chosen appropriately. Supposing that a lens combination is to be designed that is required to produce a given magnification, a number of parameters characterizing the combination can be adjusted so that the combination minimizes the defect in image formation due to a number of aberrations while, at the same time, producing the desired magnification.

For instance, a lens combination consisting of two thin convex lenses of focal lengths f_1 and f_2 can be corrected for chromatic aberration if the separation between the lenses is taken to be $d = \frac{f_1 + f_2}{2}$, provided the lenses are made of the same material. Similarly, a partial correction for spherical aberration can be effected by taking the separation to be the difference of the two focal lengths.

The entire development of the subject of ray optics has been principally geared to the understanding of the principles underlying the formation of images and those relating to the elimination, as far as possible, of the aberrations produced by optical systems.

10.14.5 Image imperfection: aberration and diffraction

In explaining the monochromatic aberrations resulting in imperfections in image formation, we have taken into account the deviations of the ray paths from the limits allowed in Gaussian optics. However, another fundamental mechanism responsible for image imperfection relates to the wave nature of light, in virtue of which there takes place a spreading and bending of wave fronts around apertures and obstacles encountered by a wave. This is referred to as *diffraction* that causes a blurring of the image of a point object formed by an optical system *regardless* of the limits set by Gaussian optics. This

diffraction effect is an irreducible one, resulting from the wave nature of light though it can be modified, within limits, by techniques of *Fourier optics* (diffraction theory will be discussed at greater length in chapter 15).

Modern day optical systems, made up of apertures, stops, lenses, and mirrors, can be of quite remarkable sophistication where aberration effects are minimized by elaborate computer-assisted designing. Such systems, however, continue to be subject to diffraction effects that result in a loss of definition of the image, and are referred to as *diffraction limited* systems.

It is to be mentioned, however, that the approach of distinguishing between diffraction effects and aberrations in an optical system is nothing but one of convenience where a number of phenomena relating to the propagation of electromagnetic waves are explained fruitfully in the ray approximation while wave aspects are taken into consideration in a higher degree of approximation. In reality, image imperfection is to be looked at as a single, complex phenomenon, of which the aberrations and the diffraction effects are nothing but two partial descriptions.

10.15 The human eye

Fig. 10-61 depicts schematically the bare essentials of the human eye, looked at as an image-forming system, omitting the intricate and exquisite anatomical details that make the system operate efficiently. In this the human eye resembles a *camera* (see sec. 10.16.1) where a lens is made to form an image of an external object on a recording screen.

Rays from the external object undergo refraction at the *cornea*, the curved anterior surface of the *eyeball*, with the *pupil* acting as an aperture stop. The function of the aperture stop is to control the amount of radiant energy entering into the eye and also to improve the definition of the final image as necessary by varying the size of the aperture.

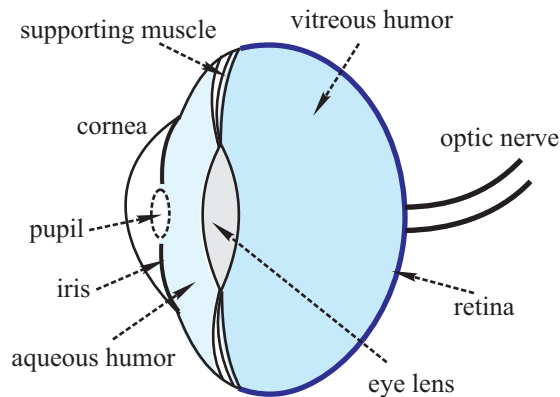


Figure 10-61: The eye as an image forming device; anatomical details are ignored; the eye lens with the two fluid media on two sides constitutes the image forming device; however, the bending of the rays as these pass through the cornea and the pupil into the aqueous humor is also of importance in image formation; the image is recorded on the retina, and sensory signals are sent to the brain by means of the optic nerve, to be converted into visual perception.

Rays are refracted at the cornea into a watery medium in the interior of the eyeball referred to as the *aqueous humor* and are then refracted by a crystalline semi-solid lens, the *eye lens*. The eye lens is held in its place by a set of muscles which also serve the purpose of effecting a change of its shape as necessary, so as to alter its focusing action. The medium on the other side of the lens, the *vitreous humor*, is thicker in comparison to the aqueous humor, though its refractive index is not much different. The rays refracted by the lens are finally focused into a real image on the *retina*, the recording screen at the back of the eye. The light focused on the retina activates a set of photo-sensors termed the *rods* and *cones* and the resulting signals are carried to the brain by the *optic nerve* for processing, so as to generate the *perception* of vision.

Considering the combination of the curved refracting surface of the cornea and the eye lens as an image forming system, the object distance u , measured from the cornea, varies from one object to another while the image distance v , i.e., the distance from the eye lens to the retina remains unchanged. In order to form a real image with this fixed value of v , the focal length of the eye-lens is made to change by altering its shape by means of the muscles supporting the eye lens - a process referred to as *accommodation*.

The maximum and minimum distances of an object that can be sharply imaged on the

retina by means of accommodation are those corresponding to points referred to as the *far point* and the *near point* respectively. For a normal eye, the far point is effectively at an infinite distance, while the distance of the near point increases with age, from around 15 cm at the age of 30 to around 40 cm at an age of 50, with considerable variability from person to person. The far point, instead of being at infinity, may also be at a large but finite distance.

A substantial change in the distance of the far point or the near point and a decrease in the power of accommodation of the eye constitute some of the *defects* of the eye, while various other defects are also possible. These defects are remedied by providing glasses to the person concerned, by replacing the eye lens with an artificial lens, or by other surgical intervention.

10.16 Optical instruments

Optical instruments can be of varied types. In particular, recent decades have witnessed a vast expansion in the practical applications of optics where a wide range of devices are used that can, in a broad sense, be called optical instruments. Thus, set-ups for the formation of holograms, various devices in fiber-optic communications, imaging devices based on optical coherence tomography (OCT), other devices for the storage and processing of optical information, optical spectrometers, and a host of other such devices of current interest can be described as optical instruments. However, in this book I will briefly touch upon the principles underlying the workings of only the *camera*, the *telescope*, and the *microscope* - the three *classical* optical instruments.

10.16.1 The camera

A camera is essentially a converging lens which forms a real image of the object to be photographed, and a recording device to capture the image, where a shutter is used to allow the passage of light through the lens for a specified duration of time. It is from this point of view that the human eye can be likened to a camera. Present day cameras are equipped with a number of sophisticated structural and functional features aimed

at improving the quality of the image and the versatility of the camera. The lens is often a combination of more than one lens elements, which is specially designed so as to be free of the aberrations and to have an appropriately large 'speed', or a correspondingly small value of the *f-number*, where the f-number of a lens is expressed by the ratio of its focal length and diameter. Lowering the f-number results in an increased brightness of the image and a smaller exposure time, which improves the image quality.

The lens and the recording device are enclosed in an appropriately designed chamber that can prevent the entry of spurious light (see fig. 10-62). Mechanical arrangements are provided for varying the separation between the camera lens and the recording screen and that between the lens elements. An adjustable diaphragm is introduced to vary the aperture of the camera lens as required.

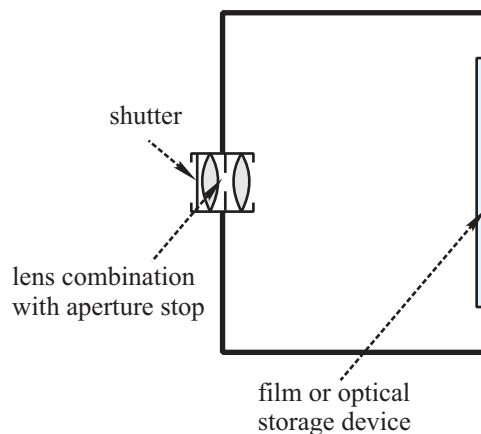


Figure 10-62: Schematic diagram to illustrate the working principle of a camera; the parameters relating to the lens combination and the aperture stop, and to the geometrical disposition of the system are so adjusted as to result in image formation with the minimum of aberrations and, at the same time, to achieve special imaging effects such as zooming and wide angle photographing; additionally, the shutter speed is important in bringing about the clarity of the image.

As I have mentioned above, camera lenses require special design because of stringent demands placed on these such as low values of the f-number, large fields of view, and the absence of aberrations. Special lens combinations such as the *Tessar lens* and the *Petzval lens* are used to meet the stringent requirements of quality image. The Tessar or the Petzval lens is made up of two lens combinations with an aperture stop placed in

between, each of the two combinations, in turn, being a doublet consisting of two lenses (fig. 10-63).

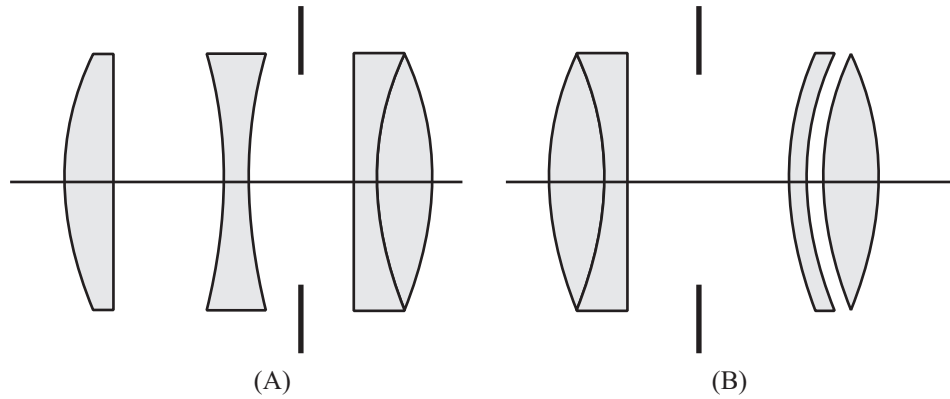


Figure 10-63: Illustrating special lens combinations for minimizing aberrations and for the achievement of a number of other special effects; (A) Tessar lens, (B) Petzval lens; each of these is made up of two combinations, with an aperture stop placed in between; each of the combinations, in turn, is a doublet, consisting of two lenses.

An important consideration relating to the functioning of the camera concerns the rate of radiant energy incident per unit area of the image and the time of exposure. The product of the two gives the total radiant energy received by the image area on the screen. The rate at which radiant energy is incident on the image area is typically of the order of $\frac{D^2}{f^2}$ where D is the effective lens diameter (the 'aperture') and f the effective focal length, and is thus inversely proportional to the square of the f-number of the camera. The exposure time, on the other hand, is related to the shutter speed. The photographer has to make a judicious choice of the two from among a set of options offered by the instrument so as to achieve maximum effect under a given set of circumstances.

The recording device in a conventional camera is a *photographic film*. While great advances have been achieved in the making of photographic films, especially in colour photography, recent decades have witnessed the development of *electronic* sensors and recording devices, resulting in the *digital camera*. The electronic recording and storage of optical information makes use of the CCD and the CMOS technologies, of which the latter is commonly used in the mobile phone cameras. These technological innovations

have brought in truly revolutionary changes in the field of photography. However, we will not concern ourselves with recording of optical information on photographic films or by electronic means here since these are not directly related to ray optics.

10.16.2 The telescope and the compound microscope

The telescope and the microscope are optical instruments designed for achieving magnification - the telescope for angular magnification and the microscope for linear magnification. The basic components of both consist of two lens combinations - an *objective* and an *eyepiece*. the objective of a telescope often consists of a single specially designed lens of large diameter and large focal length. In the *reflecting telescope* the objective is replaced with a reflecting surface of large aperture, where an auxiliary lens may also be used in conjunction with the reflector.

10.16.2.1 The telescope

Fig. 10-64 depicts the bare essentials of the optical system and of the mechanism of image formation for a refracting *astronomical* (or *Keplerian*) telescope, commonly used for the purpose of viewing heavenly bodies whose inverted images are formed by the instrument. By contrast, a *terrestrial* (or *Galilean*) telescope is used to form erect images of distant objects.

In the figure, the lens combinations of the objective and the eye-piece are depicted as single lenses for the sake of simplicity. Since a heavenly body may be assumed to be an object located at an infinitely large distance, rays originating from any point on the object to be viewed are incident on the objective in the form of a parallel bunch, making a small angle, say, α , with the axis XX' . On being refracted by the objective, which is a long-focused converging lens, these rays are focused at a point (point A in the figure) in the focal plane (the *second* focal plane, to be more precise) of the latter, A being thus the real image of the object point formed by the objective. The (second) focal plane (FF') of the objective also happens to be the first focal plane of the eye-piece, where the latter is designed to be effectively a converging lens. The rays, on converging to A, diverge thereafter, and are finally converted to a parallel bunch of rays by the eyepiece, where

these rays are inclined at an angle, say, β with the axis.

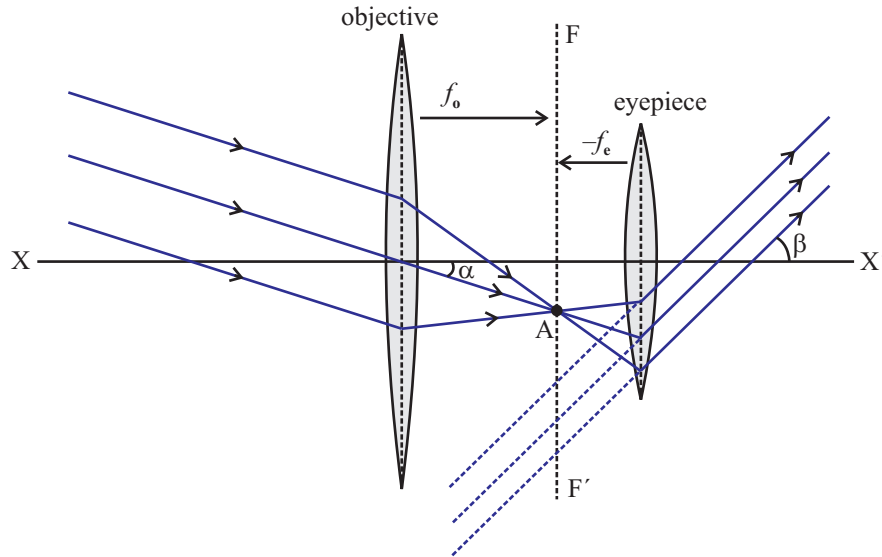


Figure 10-64: Depicting the essential features of the optical system for an astronomical telescope; FF' is the second focal plane of the objective lens, and also the first focal plane of the eye-piece; the parallel bunch of incident rays from an infinitely distant object point makes an angle α with axis of the system, while the corresponding angle for the emergent beam, which is also a parallel one, is β ; for the purpose of viewing, the eyepiece is to be moved slightly to the left so as to form a virtual image of A with a large magnification.

The telescope thereby makes possible an angular magnification

$$m_{\text{angular}} = \frac{\beta}{\alpha} = -\frac{f_o}{f_e}, \quad (10-58)$$

which can be made large in magnitude by choosing an objective with a focal length (f_o) large compared to that (f_e) of the eye-piece. For an astronomical telescope, both f_e and f_o are positive, and hence the angular magnification is negative, corresponding to the final image being an inverted one. In fig. 10-64, the final image is formed at infinity but, for convenient viewing, it can also be made to be formed as a virtual image at a large finite distance to the left of the eyepiece by shifting the latter slightly to the left of the position shown in the figure.

It follows that, in order to be effective in the observation of heavenly bodies, the objective

of a telescope has to have (a) a large aperture so as to be able to collect a relatively large amount of radiant energy coming in from the body (in addition, a large aperture minimizes diffraction effects (see chapter 15; refer also to sec. 10.14.5) that may result in a decrease of the sharpness of the image), and (b) a comparatively large focal length. In addition, it is to be specially designed to minimize the aberrations as well.

A Galilean telescope differs from an astronomical one in that the eyepiece is effectively a negative lens, for which the image of the distant object formed by the objective acts as a *virtual* object, located in its first focal plane. The angular magnification is once again given by the expression (10-58), but now it is positive (since f_o and f_e are of opposite signs), corresponding to the final image being an erect one.

Problem 10-20

The focal length of the objective of an astronomical telescope is $f_o = 1.0$ m while the telescope produces an angular magnification of $m = 50$. At what distance from the focal plane of the objective should the eyepiece be placed so that the final image formed by the instrument may be located at a distance $D = -0.5$ m from the eyepiece?

Answer to Problem 10-20

HINT: The focal length of the eyepiece is $f_e = \frac{f_o}{m} = 0.02$ m. If the distance of the intermediate image formed by the objective (in the focal plane of the latter) be u as measured from the eyepiece, then $\frac{1}{D} - \frac{1}{u} = \frac{1}{f_e}$. Substituting appropriate values, one obtains $u = -0.0192$ m. The required separation between the focal plane of the objective and the eyepiece is therefore $d = 0.0192$ m.

10.16.2.2 The compound microscope

Fig. 10-65 depicts schematically the bare essentials of the optical system of a compound microscope along with the mechanism of image formation, where the objective and the eye-piece are once again represented as single lenses for the sake of simplicity. In contrast to the telescope, the microscope objective is a short-focused one, and the angular aperture of the microscope objective is much larger. As a result, the microscope

objective is to be especially corrected for monochromatic aberrations, with particular attention to the elimination of spherical aberration and coma.

The object to be viewed, say, a small collection of biological cells, is placed on the axis (XX') of the system at a distance from the objective slightly larger than its focal length so that a magnified real image of the object is formed by the objective at the first focal plane of the eye-piece. The latter then forms a magnified final image at infinity that can be formed at a finite distance (say, the far point of the eye) as a virtual image by shifting the eye-piece slightly to the left in fig. 10-65.

The role of the eye-piece in the telescope (as also in the microscope) is to enhance the angular magnification in the following sense. If an object of height h is viewed with the bare eye by placing it at a distance, say, D then it subtends an angle $\frac{h}{D}$ at the eye. If, on the other hand, it be viewed with the help of a converging lens by placing it at the first focal plane of the latter, then the image subtends an angle $\frac{h}{f}$ (see fig. 10-66), where f stands for the focal length of the viewing lens (the 'magnifier'). Thus, by using a magnifier of short focal length, an angular magnification of $\frac{D}{f}$ can be achieved. In viewing the image formed by the objective, the eye-piece of the microscope plays the role of the magnifier. If the objective produces a linear magnification m_o , then the over-all linear magnification produced by the microscope will be

$$M \approx m_o \times \frac{D}{f_e}, \quad (10-59)$$

where the object distance for the eye-piece has been approximated by $-f_e$.

The compound microscope differs from the simple magnifier in the use of the objective as an additional component enhancing the magnification.

Problem 10-21

An object is placed at a distance $d_1 = -0.01$ m from the objective of a compound microscope, in

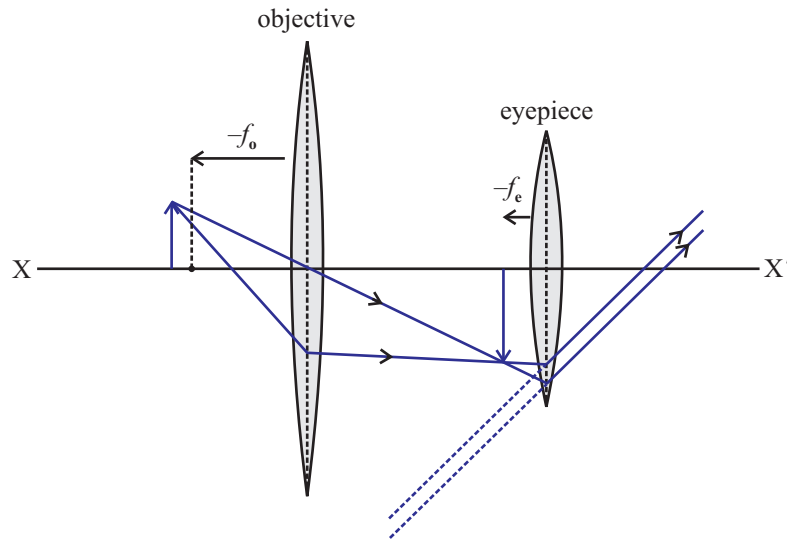


Figure 10-65: Illustrating the image formation in a compound microscope; with the object placed just beyond the first focal plane of the objective, the latter forms an intermediate image at the first focal plane of the eye-piece; the eye-piece forms a magnified final image at an appropriate distance that can be adjusted as required, for the purpose of viewing, by slightly decreasing the distance of the intermediate image from the eye-piece.

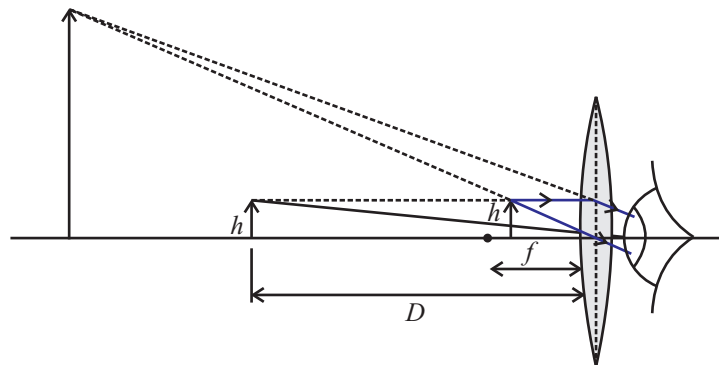


Figure 10-66: The action of a magnifier; the object is shown to be placed at a distance slightly less than the focal length so that a magnified erect image is formed; the angle subtended at the eye is $\frac{h'}{D'}$, where h is the height of the object; in the figure, the image distance is shown to be larger than the least distance of distinct vision (D); by contrast, if the object were viewed directly, without the use of the magnifying lens, the angle subtended at the eye would be $\frac{h}{D}$; in this sense, the lens effects an angular magnification of $\frac{D'}{D}$.

which the distance between the objective and the eyepiece is $L = 0.4$ m. If the intermediate real image is formed at a distance $d_2 = -0.04$ m from the eyepiece, what is the overall magnification produced by the instrument? Assume that the final image is formed at the least distance of distinct vision $D = -0.25$ m from the eyepiece and the eye. What are the focal lengths of the

objective and the eyepiece of this instrument?

Answer to Problem 10-21

HINT: The magnification produced by the objective is $m_1 = \frac{L+d_2}{d_1}$ (check this out, taking into consideration the appropriate signs of the relevant quantities), while that produced by the eyepiece is $m_2 = \frac{D}{d_2}$. Making use of the given values, the overall magnification is $m = m_1 m_2 = -225$. The focal length (f_o) of the objective satisfies $\frac{1}{f_o} = \frac{1}{L+d_2} - \frac{1}{d_1}$, i.e., $f_o = 0.0097$ m (approx), while the focal length f_e of the eyepiece is obtained from $\frac{1}{f_e} = \frac{1}{D} - \frac{1}{d_2}$, i.e., $f_e = 0.048$ m (approx).

Chapter 11

Electrostatics

11.1 Introduction: elementary charges

The elementary particles of which all matter is made up are characterized by an intrinsic property called *charge*, analogous to the other commonly known intrinsic property, namely, mass. All particles in nature can be classified into two broad groups, namely, *charged* and *uncharged* ones. Of these, the particles belonging to the first group can again be either *positively* or *negatively* charged. For instance, an *electron* is a negatively charged particle, a *proton* is positively charged, and a *neutron* is uncharged. We will see below that not only the sign of charge of a particle is of relevance, but the magnitude of charge is also a well defined physical quantity, determining the behavior of the particle under consideration in various circumstances.

If, in a body made up of such particles, the total charge due to the positively charged constituents cancels the total charge due to the negatively charged ones, then it is referred to as an *uncharged* body, while an excess of positively charged or negatively charged constituents corresponds to that body being, respectively, *positively* or *negatively* charged. A body that is either positively or negatively charged is sometimes referred to as being, simply, *charged*.

The distinction between charged and uncharged particles and bodies, and the classifi-

cation of charge into the two categories termed *positive* and *negative* charge, is based on observations relating to *forces* between these particles and bodies. A certain force is found to exist between any two charged particles or bodies. Two particles or bodies with *like* charges (the charges of both being either positive or negative) are found to repel each other, while those with *unlike* charges (one being of positive charge and the other of negative charge) attract.

1. Under some circumstances a pair of bodies with like charges are found to *attract* each other. This is due to the effect of electrical *induction* (see sec. 11.3.1).
2. Referring to the two opposite types of charges, it is a matter of convention as to which to call positive and which negative. Instead of the presently adopted convention, one could follow a different convention where an electron would be characterized by a positive charge and proton by a negative one. In that case the signs of charges of all charged bodies would be opposite to those according to the present convention.

A number of observations and measurements have shown that there exists a certain *minimum* and *indivisible* unit of charge, this unit being the charge of the proton and the electron for positive and negative charge respectively while, moreover, the *magnitudes* of these two are *equal*. In other words, the charge of an electron is equal and opposite to that of a proton. The charge of a body depends on how many electrons and protons are there in it. If these numbers are respectively N_- and N_+ , and if the magnitude of the minimum charge referred to above be e (i.e., the charge of the electron and the proton be respectively $-e$ and $+e$), then the charge of that body will be given by

$$q = (N_+ - N_-)e. \quad (11-1)$$

Since, in this formula, N_+ and N_- are positive integers, the charge q of a body is necessarily a positive or negative integral multiple of the elementary charge e , while the value $q = 0$ corresponds to N_+ and N_- being equal to each other. It follows that the charge of a body can have only a *discrete* set of values. However, for a *macroscopic* body, the numbers N_+ , N_- , and $N_+ - N_-$ are, in general, so large in magnitude that the charge

q can effectively be considered to be a quantity that can be made to vary continuously. Moreover, in practice, N_+ and N_- are not well defined numbers for a macroscopic body since these are subject to *fluctuations*, and only their *average* values are meaningful.

In the SI system of units, the unit of charge is defined with reference to the unit of current, namely the *ampere* (A), and is named the *coulomb* (C). Expressed in this unit, the magnitude of the elementary charge introduced above is $e = 1.6 \times 10^{-19}$ C. In other words, the charge of an electron and a proton are respectively -1.6×10^{-19} C and 1.6×10^{-19} C (approx).

11.2 Acquisition of charges by bodies

11.2.1 Transfer of elementary charges

The question of which particles are to be considered as the ultimate or elementary constituents of matter, is a complex one. For our present purpose we consider protons, neutrons, and electrons as the elementary constituents. Protons and neutrons bind to one another, forming nuclei. Electrons are external to the nuclei, being bound to these by the attractive force between opposite charges. The number of electrons so attached to a nucleus is equal to the number of protons in it, as a result of which an uncharged atom is formed. A molecule is usually formed of a number of atoms bound to one another, while a large number of molecules, bound together by relatively weak cohesive forces make up a macroscopic body (for a gaseous substance, the cohesive force is negligible).

Strictly speaking, the protons and neutrons do not count as elementary constituents of matter, because these in turn are made up of constituent particles called *quarks*. Quarks can be of various different types, among which two specific types of quarks, together with their *anti-particles*, i.e., particles with complementary sets of properties, make up the protons and neutrons. The charges of these quarks are not positive or negative integral multiples of the elementary unit of charge e , but are *fractional* multiples thereof. However, these quarks and anti-quarks are not found in isolation since they always remain bound to one another, the charges of the resulting composite

particles being integral multiples of the elementary charge e .

At times, an electron in an atom or a molecule acquires energy from external sources and is freed from it. In certain circumstances such electrons detached from the bondage of the atoms or molecules are transferred from one body to another.

When one or more electrons are freed from an atom or a molecule then that atom or molecule becomes positively charged due to a deficiency of negatively charged electrons. Such a charged atom or molecule is called an *ion*. Ions can be produced by other means as well resulting, under certain circumstances, in even negatively charged ions formed by the addition of electrons to neutral atoms.

If the electrons detached from the atoms or molecules of a body remain in it without being transferred to any other body, then they may remain associated with large collections of molecules in the body and the body as a whole remains neutral since the number of positively charged and negatively charged constituents in it continue to be equal ($N_+ = N_-$).

In other words, one can talk of two types of electrons in a material - those bound relatively strongly to individual atoms or molecules, and the ones that are not bound to specific atoms or molecules, moving relatively freely through the material, being bound to large collections of atoms or ions as a whole.

If, by some means, a number of bound or free electrons in a body are transferred to some other body, then there occurs a relative deficiency of the negatively charged electrons in the first body ($N_- < N_+$) which thereby becomes positively charged. The body acquiring these electrons, on the other hand, becomes negatively charged, with a relative excess of electrons ($N_- > N_+$). However, there exist processes other than the transfer of electrons as well whereby a body can acquire a charge. For instance, ions or other charged particles floating in the atmosphere can get attached to a body, developing a charge in it.

We have seen in chapter 8 that even when a body is at rest, its microscopic constituents

continue to be in incessant random motion. The temperature of the body is an index of such random motion of these constituents, which is therefore referred to as *thermal* motion. However, such thermal motion notwithstanding, the *average* velocity of the microscopic constituents, including that of the charged ones, turns out to be zero. If the charged constituents do not acquire any additional velocity due to any other cause, then one can say that the thermal motion does not result in any *flow* of the charges. The charge in a body is then referred to as *static* charge. For instance, the charge in an isolated body held at rest is an instance of static charge.

In contrast, if the two ends of a wire be connected to the two terminals of an *electrical cell*, then a *drift motion* of the free electrons in the wire is produced (see chapter 12 for an introduction to the relevant concepts), resulting in a *current* being set up in the wire.

The study of effects resulting from static charges in bodies constitutes the subject of *electrostatics*, which we look at in this chapter. Phenomena relating to the flow of electrical currents will be taken up in chapters 12 and 13.

A charged stationary particle or body creates a static *electric field* in the region of space around it, while *magnetic fields* result from currents set up in materials. I will introduce these concepts in the present chapter and in chapter 12.

11.2.2 Contact electrification and contact potential

Suppose A and B are two bodies made of different materials, kept in contact with each other, with a common surface of contact between them as in fig. 11-1. It is usually found in such a situation that electrons are transferred from one of the bodies to the other, as a result of which one of the bodies becomes charged positively, and the other negatively. The process of developing electrical charges in bodies by bringing them in contact is referred to as contact electrification. In this process a large *electrical potential* difference is also developed between the contiguous parts of the two bodies in contact (see sec. 11.4.2 for an introduction to the concept of electrical potential), and is known as *contact potential*.

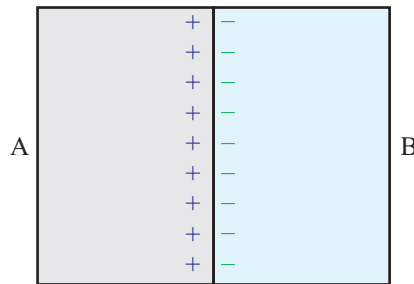


Figure 11-1: Electrification by contact; electrons are transferred from the body marked A to the one marked B; the former thereby acquires positive charge and the latter negative.

Contact potentials are found to play a significant role in various phenomena of our daily experience, in numerous natural phenomena, and in a considerable number of technical appliances. For instance, the action of an electrolytic cell or that of a semiconductor diode depends on the generation of a contact potential.

A contact potential is not necessarily developed by the transfer of electrons alone.

In an electrolytic cell, for instance, a contact potential is developed not only by the transfer of electrons, but by the transfer of ions as well.

Electrification of bodies is often effected by *rubbing* together bodies made of dissimilar materials, a process commonly referred to as electrification by friction. This, however, is not a process distinct from contact electrification, because the charges are developed in this process essentially due to contact between dissimilar materials. The role of rubbing or friction here is simply to make a better contact and to increase the effective contact area.

The explanation of contact electrification from fundamental principles, however, is not known with adequate clarity, nor does a unified theory seem likely. For instance, the mechanism of contact electrification for a pair of insulators differs from that for a pair of metals. A detailed explanation in either case requires that *quantum theory* be invoked. We will, however, not enter here into a discussion of the origins of contact electricity.

11.3 Electrostatic force between charges

11.3.1 Coulomb's law

As I have mentioned, the observational basis for deciding whether a body is charged and, if so, determining what the sign and magnitude of the charge is, is provided by a certain type of force, namely the force arising between bodies by virtue of their charges. The basic principle that gives us the magnitude and direction of force between two charged bodies is *Coulomb's law*.

There may arise an electrical force between a charged body and one that is uncharged, once again because of electrical induction (see below). Even though a body may possess zero *net* charge, parts of it may contain positive charge, with an equal and opposite charge developing in other parts. The resultant of the forces between a charged body and these various parts of an uncharged body may be non-zero. More precisely, it is not the net charges of bodies, but their charge *distributions* that determines the forces of electrical origin between bodies. Knowing the charge distributions, one can work out the force between two bodies by making use of Coulomb's law, which gives the force between two charged *particles*.

Suppose that there are two charged particles in an evacuated region of space, which we label as '1' and '2', with charges q_1 and q_2 respectively. If the separation between the particles be r then the magnitude of the electrical force between them is given by

$$F = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r^2}, \quad (11-2)$$

where q_1 and q_2 are to be interpreted as the magnitudes of the charges carried by the two particles. Here ϵ_0 is a constant, referred to as the *permittivity of free space*. Its value in the SI system is $8.85 \times 10^{-12} \text{ C}^2 \cdot \text{N}^{-1} \cdot \text{m}^{-2}$. As for the direction of the force, it acts along the line joining the two particles and is a repulsive one for two *like* charges, i.e., charges with the same sign. For *unlike* charges, on the other hand, the force is attractive.

These facts relating to Coulomb's law can all be combined into a single equation involv-

ing vectors. Let \mathbf{r}_1 and \mathbf{r}_2 be the position vectors of the two charged particles relative to any chosen origin and $\mathbf{r}_{12} = \mathbf{r}_1 - \mathbf{r}_2$ denote the position vector of the particle '1' relative to particle '2' (similarly, $\mathbf{r}_{21} = \mathbf{r}_2 - \mathbf{r}_1$ stands for the position vector of '2' relative to '1'). Then the force \mathbf{F}_{12} on '1' exerted by '2' by virtue of the charges on these two particles is given by the expression

$$\mathbf{F}_{12} = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r_{12}^3} \mathbf{r}_{12}, \quad (11-3a)$$

where q_1 and q_2 are the charges of the two particles, with appropriate signs. Here $r_{12} = |\mathbf{r}_1 - \mathbf{r}_2|$ is the magnitude of the distance between the two particles, which can also be written as r_{21} .

The force \mathbf{F}_{21} exerted by '1' on '2' can be expressed in a similar manner and satisfies (see fig. 11-2)

$$\mathbf{F}_{21} = -\mathbf{F}_{12}. \quad (11-3b)$$

This shows that the forces exerted by the two charged particles on each other are equal and opposite, conforming to Newton's third law and is, moreover, *central* in nature (see sec. 3.17.2).

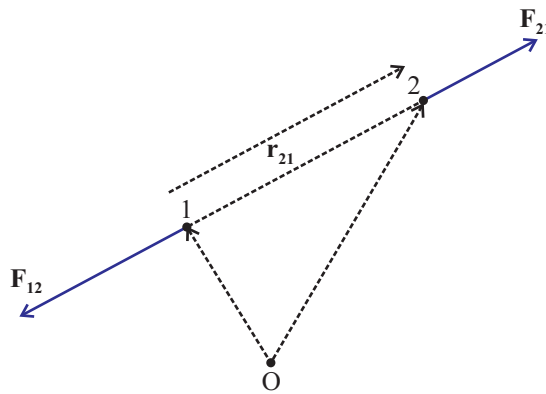


Figure 11-2: Illustrating Coulomb's law of force between two charged particles labeled '1' and '2'; the force between the particles is central in nature, i.e., acts along the line joining the particles; it can be repulsive or attractive, depending on the signs of the two charges; the position vectors of the two charge-points relative to a chosen origin O are respectively \mathbf{r}_1 and \mathbf{r}_2 , while \mathbf{r}_{21} is the position vector of '2' relative to '1'; \mathbf{F}_{12} and \mathbf{F}_{21} are forces on '1' and '2' respectively.

The force law (11-3a) resembles the law of gravitational interaction between two point masses (see sec. 5.1). In both these two cases, the force varies inversely as the square of the distance between the interacting particles and acts along the line joining the two. These features correspond to what are known as *inverse square central force* fields. The difference between the two interactions lies in the fact that the gravitational force exists between all pairs of point masses and is always attractive, while the electrostatic force operates only between charged particles and may be attractive or repulsive. Another major point of distinction relates to the *medium* in which the particles causing the forces are situated. The formula (5-3a) describing the gravitational force between two particles holds regardless of the medium in which the particles are located, while the formula (11-3a) holds only for two particles located in an evacuated region (i.e., in free space). For a pair of charged particles located in any material medium the description of the force between the two involves further considerations. In the case of a *dielectric* medium, for instance (refer to sec. 11.10), the force formula assumes a form similar to (11-3a), but with the constant ϵ_0 replaced with some other constant characteristic of that medium. For the present, however, we confine all our considerations to static charges located in free space.

The similarity between the force laws in gravitational and electrostatic interactions implies a correspondence between the concepts and mathematical relations developed in the theory of gravitation and those of electrostatics. For instance, the concepts of the electrical field intensity and potential (see below) are closely analogous to those of the gravitational intensity and potential and, similarly, Gauss' principle and its consequences in gravitation are closely resembled by corresponding results in electrostatics. Thus, many of the concepts and derivations presented in this chapter will be similar to corresponding concepts and derivations in chapter 5.

Digression: Electrical induction.

Electrical induction is the process of charge separation in a body due to the electrical influence of some other charged body placed in its vicinity, where the two bodies are not brought into contact. Imagine, for instance, a positively charged body A, say, a

conductor, brought near an uncharged body B, where the latter may be a conductor or a dielectric (see sec. 11.10 for an introduction to conductors and dielectrics). The charge on A exerts electrostatic forces on the charges in B, where these charges are *free* ones in the case of a conductor and *bound* ones in the case of a dielectric, causing a charge separation in B (more generally, more than one charged bodies may cause the charge separation in B).

This charge separation appears principally in the form of *surface charges* on the boundary surface of B, with a preponderance of negative charges in regions relatively close to A, and of positive charges in regions away from A. However, the mechanisms underlying the charge separation in the case of a conductor differs from that in the case of a dielectric, i.e., an insulator. In the case of B being a conductor, the charge separation arises by way of motion of the free electrons in B. In the case of a dielectric, on the other hand, the relevant mechanism is that of *polarization* in the dielectric, where a *volume* charge distribution may appear in addition to the surface charges.

We now revert to a consideration of charges and their interactions in free space, till we turn our attention to electrostatics of material media in sec. 11.10 (and in some parts of sec. 11.9).

11.3.2 The principle of superposition

One other basic principle of electrostatics, once again similar to the corresponding principle in gravitation, is the *principle of superposition*. Suppose that, instead of just two particles, we have *three* charged particles labeled as say, '1', '2', and '3', interacting with one another. What will then be the force experienced by any one of these, say '1', due to the other two, i.e., due to '2' and '3' ? The principle of superposition states that this force, which we denote by F_1 in the present instance, is the sum of two terms,

$$F_1 = F_{12} + F_{13}, \quad (11-4a)$$

where F_{12} and F_{13} are the forces exerted on '1' by '2' and '3' respectively in accordance with Coulomb's law stated above. Thus, if r_1 , r_2 , and r_3 be the position vectors of the

three particles with respect to any chosen origin and q_1, q_2, q_3 their respective charges, then

$$\mathbf{F}_1 = \frac{1}{4\pi\epsilon_0} \left[\frac{q_1 q_2}{|\mathbf{r}_1 - \mathbf{r}_2|^3} (\mathbf{r}_1 - \mathbf{r}_2) + \frac{q_1 q_3}{|\mathbf{r}_1 - \mathbf{r}_3|^3} (\mathbf{r}_1 - \mathbf{r}_3) \right]. \quad (11-4b)$$

More generally, for a system of N charged particles with charges q_1, q_2, \dots, q_N at locations given by position vectors $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N$ with respect to any chosen origin, the force of electrostatic interaction on any one particle, say the i th one ($i = 1, 2, \dots, N$) due to all the others is given by the expression

$$\mathbf{F}_i = \sum_{j(j \neq i)} \frac{1}{4\pi\epsilon_0} \left[\frac{q_i q_j}{|\mathbf{r}_i - \mathbf{r}_j|^3} (\mathbf{r}_i - \mathbf{r}_j) \right], \quad (11-4c)$$

where the summation is to be carried out for all j from 1 to N , excluding the value $j = i$.

The force between two charged particles depends on their charges. How, then, are the charges defined in the first place? In reality, one defines the charges in terms of the force between them. This seems to be a logical paradox. However, all one needs is a *consistent* scheme where one has a well-defined operational procedure of measuring the charges *and* a prescription for predicting the force between two particles whose charges have been determined in accordance with that procedure. The consistency of the scheme then means that the forces so predicted should agree with the rates of change of momenta of the respective particles. Such a consistent scheme is provided by Coulomb's law as given by eq. (11-3a) and the principle of superposition taken together, along with the laws of mechanics.

One can determine the force exerted by one extended charged body on another in a similar manner. Thus, considering any one of the two bodies as a collection of charged particles, one can work out the force on any one of those particles exerted by all the charged particles making up the other body as in eq. (11-4c), and then take the vector sum of the forces on all the charged particles in the first body. The force on the second body can also be worked out similarly. In addition, such an approach gives the *torques*

on the two bodies as well.

Starting from Coulomb's law and the principle of superposition, I will now introduce the concepts relating to the electric field of a system of charges, the electric field intensity at any specified point, and of electrical potential.

11.4 Electric field intensity and potential

11.4.1 Electric field intensity

Fig. 11-3 depicts a point charge q located at a point P, due to which a force is exerted on another point charge q' located at R. Since this second charge q' and the point R can be chosen arbitrarily, one can say that the charge q located at P creates a certain *influence* around itself by virtue of which it exerts a force on a second charge located at any point such as R (we refer to this as the *Coulomb force*; the force on the charge q exerted by the second charge q' is not relevant in the present context).

If a *standard* or *reference* charge is placed at any such point R, then the force on that charge can be taken as a quantitative measure of the influence at R set up by the charge q . If this reference charge at the point R is taken to be unity ($q' = 1$ C in the SI system), then the force on it exerted by the charge q is referred to as the electric field intensity ('electric intensity' in brief) at R due to the charge q at P. In other words, the intensity is the force on a unit charge placed at the point under consideration. Evidently, the electric field intensity is a vector quantity, which can be obtained in the present instance from eq. (11-3a) on putting $q_1 = 1$ C, $q_2 = q$, and, say, $\mathbf{r}_{12} = \mathbf{r}$, the position vector of the point R relative to P. Denoting the intensity by \mathbf{E} one obtains

$$\mathbf{E} = \frac{q}{4\pi\epsilon_0 r^3} \mathbf{r}. \quad (11-5a)$$

Alternatively, assuming that the charge q (referred to as the *source* charge) is located at a point P with position vector \mathbf{r} (the *source* point) relative to a chosen origin, the electric intensity at a second point R (referred to as the *field* point) with position vector \mathbf{r}' is

given by the expression

$$\mathbf{E} = \frac{q}{4\pi\epsilon_0} \frac{\mathbf{r}' - \mathbf{r}}{|\mathbf{r}' - \mathbf{r}|^3}. \quad (11-5b)$$

It helps to write this in the simpler form

$$\mathbf{E} = \frac{q}{4\pi\epsilon_0 u^2} \hat{u}, \quad (11-5c)$$

where u stands for the distance from the source point to the field point,

$$u = |\mathbf{r}' - \mathbf{r}|, \quad (11-5d)$$

and \hat{u} is the unit vector directed from the former to the latter,

$$\hat{u} = \frac{\mathbf{r}' - \mathbf{r}}{|\mathbf{r}' - \mathbf{r}|}. \quad (11-5e)$$

The electric field intensities at various different points like R due to the source charge q located at P, make up a *field* of intensities or, equivalently, an *electric field*. It is the electric field that describes completely the influence set up by the charge q at various points in space.

The term source points and field points are used generally to distinguish between the points where a given set of source charges (i.e., the charges creating an electric field) are located and those at which one wishes to obtain the field intensities due to these source charges.

Knowing the intensity \mathbf{E} at a field point R, the force on a charge, say q' placed at the field point can be obtained as (recall the definition of intensity as the force on a unit charge) $\mathbf{F} = q'\mathbf{E}$.

Now imagine a number of source charges q_1, q_2, \dots, q_N located at points with position vectors, say, $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N$ respectively, instead of a single source charge q . Together,

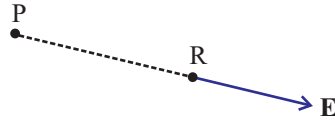


Figure 11-3: Illustrating the concept of electric field intensity; a charge q at the source point P generates a field around it, the intensity at the field point R being \mathbf{E} .

all these charges set up an electric field in space, which is described in terms of the electric field intensities at various field points. Considering a typical field point, say \mathbf{r} , the intensity at this point is defined as the force on a unit charge imagined to be placed at \mathbf{r} . Making use of Coulomb's law and the principle of superposition as expressed by eq. (11-4c), the intensity is seen to be

$$\mathbf{E} = \frac{1}{4\pi\epsilon_0} \sum_{i=1}^N \frac{q_i(\mathbf{r} - \mathbf{r}_i)}{|\mathbf{r} - \mathbf{r}_i|^3}. \quad (11-6a)$$

Correspondingly, the force on a charge q placed at the field point is given by

$$\mathbf{F} = q\mathbf{E}. \quad (11-6b)$$

(check the above statements out).

Fig. 11-4 depicts schematically an electric field in a region R of space, where a vector \mathbf{E} , represented by a directed line segment is associated with every point \mathbf{r} in the region. One thereby has a vector function of the vector variable \mathbf{r} which can be denoted by the symbol $\mathbf{E}(\mathbf{r})$. This vector function is given by an expression of the form (11-6a) for a field set up by a number of point charges. This is a particular instance of a *vector field* introduced in sec. 2.13, another instance of which is provided by a gravitational field discussed in chapter 5. Both an electric field and a gravitational field are instances of a *field of force* (see sec. 3.11). Indeed, considering any given charge, say, q , a field of force $\mathbf{F}(\mathbf{r})$ can be obtained from the electric field $\mathbf{E}(\mathbf{r})$ by using eq. (11-6b), this being the force experienced by the charge q at various points in the electric field, and the electric field is simply the field of force for $q = 1$.

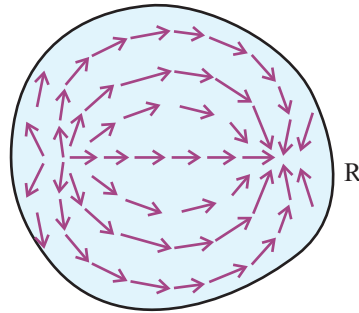


Figure 11-4: An electric field (schematic) set up in a region R (source charges not shown in the figure), represented by an intensity vector at every point in the region.

11.4.2 Electrical potential

An important feature of the force field $\mathbf{F}(\mathbf{r}) = q\mathbf{E}(\mathbf{r})$ mentioned in sec. 11.4.1 is that it is a *conservative* one.

Recall from sec. 3.15.2 that a force field acting on a particle is said to be conservative if, given any two points P and Q in the field, the work done by the force in a displacement from Q to P along any given path turns out to be independent of the path, being determined solely by the positions of the two points P and Q . An equivalent way of stating this is to say that the line integral $\int_Q^P \mathbf{F} \cdot d\mathbf{r}$ evaluated along a path connecting P and Q is independent of the path chosen. Recall further that, for a conservative force field, one can define a potential energy (V) of the particle where the difference of potential energies at P and Q (i.e., $V_P - V_Q$) is the work done *against* the force in a displacement from Q to P .

This means that the work done by the electric field in a displacement of a particle of charge q from any point Q to another point P along an arbitrarily chosen path is independent of the path followed. Further, a potential energy of the particle at any chosen point, say, P with position vector \mathbf{r} , can be defined as the work done against the field in a displacement of the particle from a chosen reference point to the point P under consideration. If now the charge q of the particle is chosen to be unity (1 C in the SI system) then the potential energy of the charge is referred to as the electrical *potential*

(or, in brief, the potential) at the point \mathbf{r} . With $q = 1$, the force $\mathbf{F}(\mathbf{r})$ reduces to the electric field intensity $\mathbf{E}(\mathbf{r})$ and thus the potential $V(\mathbf{r})$ is given by the formula

$$V(\mathbf{r}) = - \int_{\mathbf{r}_0}^{\mathbf{r}} \mathbf{E} \cdot d\mathbf{r}, \quad (11-7)$$

where \mathbf{r}_0 is the position vector of the chosen reference point and the integration is performed along any arbitrarily chosen path from the reference point to the point under consideration.

The potential so defined is undetermined to the extent of an additive constant since a different choice of the reference point causes a change in the value of $V(\mathbf{r})$ by a constant amount (check this out).

The electrical potential V being the potential energy of a unit charge at any given point in an electric field, the potential energy of a charge q is given by the expression

$$\mathcal{V} = qV, \quad (11-8a)$$

and the work done *against* the field in displacing the charge q from a point \mathcal{Q} to another point \mathcal{P} is

$$W_{\mathcal{Q} \rightarrow \mathcal{P}} = q(V_{\mathcal{P}} - V_{\mathcal{Q}}). \quad (11-8b)$$

Incidentally, for a *uniform* electric field of magnitude E directed along, say, the x-axis of a co-ordinate system, the potential difference between any two points \mathcal{Q} and \mathcal{P} separated by a distance d (with co-ordinates, say, x_0 and $x_0 + d$ respectively) is given by the formula

$$V_{\mathcal{P}} - V_{\mathcal{Q}} = -Ed. \quad (11-9)$$

This can be established straightaway by choosing the reference point on the x-axis at, say, the origin \mathcal{O} and applying the formula (11-7) twice, once for \mathcal{O} and \mathcal{P} , and then for \mathcal{O}

and Q , integrating along the x -axis in both the cases, and finally subtracting one result from the other (check this out).

Since the unit of energy in the SI system is the joule (J), the unit of potential will be $\text{J}\cdot\text{C}^{-1}$. This unit is referred to as the *volt* (V).

11.4.3 Electrical potential: summary

I now summarize for you what I have said above regarding electric field intensity and potential in an electric field.

1. The field of force acting on a charged particle in an electric field created by stationary charges is conservative in nature.
2. One can define a potential at any point in such a field with respect to a chosen reference point. If the position vector of the point be \mathbf{r} , then the potential $V(\mathbf{r})$ is the work done against the force due to the field in transferring a unit charge from the reference point to the point \mathbf{r} . The unit of electrical potential is the volt (V).
3. The potential energy of a charge q placed at the point \mathbf{r} is $qV(\mathbf{r})$.
4. If P and Q be any two points in an electric field, the work done against the electrical force in transferring a charge q from Q to P along any chosen path is given by the expression (11-8b).
5. If a different reference point is chosen in defining the potential then the latter gets changed by the addition of a constant term, but this non-uniqueness of the potential does not show up in the potential *difference* between any two points since the constant additive term gets canceled in the potential difference.
6. For a given choice of the reference point, the potential *at* that point will evidently be zero.

11.4.4 Potential 'at infinity'

The potential at any point in an electric field depends on the choice of the reference point through an additive constant while, for a given choice of the reference point the

potential *at* that reference point is zero. In numerous situations of interest the reference point is taken to be a point at infinity.

The phrase ‘point at infinity’ needs clarification. Starting from any point, say, O (see fig. 11-5), imagine two lines OA and OB extending along two different directions. Assuming that the reference point is located at a distance d from O on any one of these lines and considering the limit $d \rightarrow \infty$, one approaches the ‘point at infinity’ along that direction. One thus gets two such ‘points at infinity’ for the two directions chosen. One could equally well have chosen a direction other than the above two and arrived at another ‘point at infinity’. If all these possible choices of the reference point at infinity be equivalent, then one can talk of a ‘point at infinity’ regardless of the direction along which that point is approached. Else, the direction along which a point at infinity is approached will have to be taken into account.

Considering any two given directions, say, along OA and OB in fig. 11-5, the point at infinity approached for $d \rightarrow \infty$ along any one of these will be equivalent to the point at infinity approached along the other, if the work done by (or against) the electrical force field in displacing a charged particle from one of these points at infinity to the other along any path (dashed line in the figure) be zero.

Under what conditions will the above requirement be met with for all pairs of lines like OA and OB ? Suppose that all the source charges responsible for the setting up of the field are located within a finite distance from O (which can be taken as the origin in the present context), i.e., all these source charges are situated within a finite region of space. This then will guarantee that all ‘points at infinity’ are equivalent as reference points. Consequently, any one of these can be referred to as *the* point at infinity, and the potential at such a point may then be taken to be zero.

If, on the other hand, some of the source charges setting up the field are spread out to infinite distances, then all the various points at infinite distances approached along different directions will not be equivalent as reference points and one will have to specify a particular direction along which the reference point at infinity will have to be ap-

proached. In some situations of this type, one chooses a point situated at a *finite* distance from a chosen origin as the reference point.

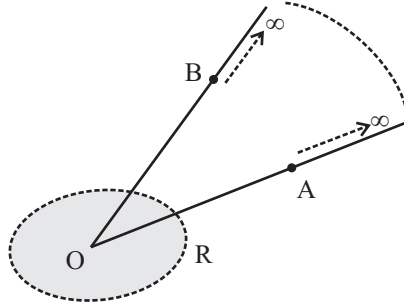


Figure 11-5: Explaining the idea of potential at infinity; imagine points located on OA and OB at infinite distances from any chosen point O; the dotted curve is a path joining these two 'points at infinity'; the work done by the electric field in displacing a charge along this path has to be zero; if this requirement is met with for all pairs of lines like OA and OB, then all such 'points at infinity' are equivalent; this, in turn, requires that all the source charges be contained within a finite region, say, R; one can then take the point at infinity as the reference point in defining the potential.

In the case of a *continuously* distributed system of charges, one can define a charge *density* at any point as the ratio

$$\rho = \frac{\delta q}{\delta V}, \quad (11-10)$$

of the charge δq contained within an infinitesimal volume δV around that point. In this case the charge density should go to zero sufficiently rapidly along any and every line (such as OA or OB in the figure) in order that a 'point at infinity' may make sense and that one can choose the potential to be zero at that point.

Problem 11-1

The electrical potentials of three large conducting bodies (A, B, C) are $V_1 = -10$ V, $V_2 = 50$ V, and $V_3 = 100$ V respectively where the potential at infinity is taken to be zero. An amount $q_1 = 1.0 \times 10^{-10}$ C is transferred from A to B, and $q_2 = -2.0 \times 10^{-10}$ C from B to C. Assuming that the potentials of A, B, and C do not change appreciably due to these charge transfers, work out the change in the potential energy of the system.

Answer to Problem 11-1

HINT: Since the potential energy of a charge q located at a point at potential V is qV , the change in potential energy resulting from a transfer of charge q_1 through a potential difference $V_2 - V_1$ is $q_1(V_2 - V_1)$ (it does not matter that the charge is not concentrated at one point since the surface of a conductor is an equipotential one - see sec. 11.10.3), and that due to a transfer of charge q_2 through a potential difference $V_3 - V_2$ is $q_2(V_3 - V_2)$. In other words, the total change in potential energy is $(1.0 \times 60 + (-2.0) \times 50) \times 10^{-10}$ J, i.e., -4.0×10^{-9} J.

Problem 11-2

The electric field intensity between two large parallel conducting plates can be assumed to be uniform, being in a direction perpendicular to the plates (this constitutes a parallel plate capacitor, see sec. 11.11.8). If the force on an electron placed at any point between the plates be 3.2×10^{-16} N, and if the separation between the plates is 0.01 m, calculate the field strength, and the potential difference between the plates.

Answer to Problem 11-2

HINT: The force on the electron is $F = qE$, where $q = 1.6 \times 10^{-19}$ C, and E is the field strength (we consider magnitudes only, keeping the directions implied). With the given value of F , we get $E = 2000 \text{ V}\cdot\text{m}^{-1}$. If the separation between the plates be d , then the potential difference V between them is related to E as $E = \frac{V}{d}$ (refer to eq. (11-9); recall that we are working with magnitudes only; the intensity is directed from the higher to the lower potential), from which one gets $V = 20$ V.

11.4.5 Potential due to a point charge

Usually, if all source charges are located within a finite distance from the origin, the reference point is taken to be at infinity in the sense discussed above, i.e., the potential at infinity is assumed to be zero (in this book we will consider mostly such situations). For instance, consider first the simplest of situations, where there is only *one* source charge located at a point, say, P with position vector \mathbf{r} relative to a chosen origin O (see fig. 11-6). What will be the potential due to this charge at a field point, say, R, with

position vector \mathbf{r}' ? By definition, the potential at R is the work done against the force exerted by the source charge (say, q) when a unit charge is brought to the position R from an infinite distance along any chosen line. In the present instance it is convenient to choose this as the line joining P to R, imagined to be extended to infinity. One can then work out the expression for the work done against the force exerted on the unit charge by the source charge q , which turns out to be

$$V = \frac{1}{4\pi\epsilon_0} \frac{q}{|\mathbf{r}' - \mathbf{r}|}, \quad (11-11a)$$

or, more simply,

$$V = \frac{q}{4\pi\epsilon_0 u}. \quad (11-11b)$$

In this last expression, u stands for the distance from the source point to the field point, as in eq. (11-5d).

Problem 11-3

Check the above statements out, leading to equations (11-11a), (11-11b).

Answer to Problem 11-3

HINT: Denoting by u' the distance from P to R' (see fig. 11-6), the work done against the force exerted by the source charge in displacing a unit charge from a distance u' to $u' + \delta u'$ is $-\mathbf{F}(u') \cdot \hat{u} \delta u'$, where, according to eq. (11-5c), $\mathbf{F}(u') = \frac{q}{4\pi\epsilon_0 u'^2} \hat{u}$. Here \hat{u} is the unit vector along the vector $\mathbf{u} = \mathbf{r}' - \mathbf{r}$. Breaking up the path (along which the unit charge is transferred) into a large number of small segments and summing up for all these segments one gets the required potential. In the limit of the lengths of the segments going to zero, the required expression reduces to the integral

$$V = -\frac{q}{4\pi\epsilon_0} \int_{\infty}^u \frac{du'}{u'^2}. \quad (11-12)$$

.

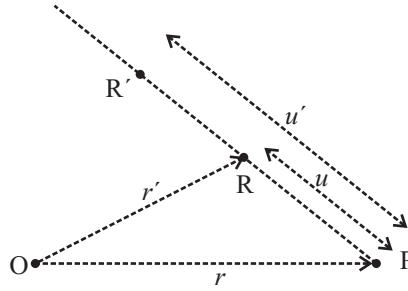


Figure 11-6: Potential due to a point charge; a source charge q at P produces a potential V at the field point R; imagining the line PR to be extended out to infinity, the potential is the work done against the force exerted by the source charge in bringing a unit charge from infinite distance down to R along this line; R' is an intermediate point arrived at by the unit charge in this process; O is any chosen origin.

11.4.6 Potential due to a number of point charges

We will now obtain an expression for the potential at any point in the field set up by a number of source charges. Let N number of point charges q_1, q_2, \dots, q_N be located at points with position vectors $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N$ with respect to any chosen origin. According to the principle of superposition, the potential V at a field point with position vector \mathbf{r} will be the sum of terms V_i ($i = 1, 2, \dots, N$), where V_i is the potential at the field point due to the charge q_i located at \mathbf{r}_i independently of the other source charges (reason this out). Using for each source charge an expression of the form (11-11a), one gets

$$V_i = \frac{1}{4\pi\epsilon_0} \frac{q_i}{|\mathbf{r} - \mathbf{r}_i|}, \quad (11-13a)$$

and

$$V = \sum_{i=1}^N V_i = \frac{1}{4\pi\epsilon_0} \sum_{i=1}^N \frac{q_i}{|\mathbf{r} - \mathbf{r}_i|}. \quad (11-13b)$$

The following expression looks simpler:

$$V = \frac{1}{4\pi\epsilon_0} \sum_{i=1}^N \frac{q_i}{u_i}. \quad (11-13c)$$

In this last expression u_i stands for the distance from the i th source point to the field point.

Note that the equations (11-6a) and (11-13b) are obtained from (11-5b) and (11-13a) respectively by invoking the superposition principle. However, while eq. (11-6a) involves a *vector* sum of contributions from individual source charges, eq. (11-13b) gives the potential as a sum of *scalar* contributions. As a result, the calculation of the potential at a point in an electric field is sometimes more convenient than that of the intensity. However, it is the intensity that is of more direct physical relevance, and one has to work out the *rates of change* of the potential with distance along various directions so as to arrive at the intensity.

Problem 11-4

Charges $-q$, q , $-q$, q are placed at the four corners A, B, C, D, of a square of side a , so that the charges at the two ends of any side are of opposite signs while those at the ends of any diagonal are of the same sign. Find the energy required to set up this charge distribution.

Answer to Problem 11-4

HINT: Imagine the charge $-q$ placed at the corner A all by itself, without the other three charges being present. If one now brings in the charge q from infinity up to B without allowing the charge at A to move, then one has to perform work $W_1 = -\frac{q^2}{4\pi\epsilon_0 a}$, where we make use of the fact that the potential due to the charge $-q$ (located at A) at the point B is $V_1 = -\frac{q}{4\pi\epsilon_0 a}$, and that the potential energy of the charge q located at B (in the absence of charges at C and D) is qV_1 . Similarly, if one now brings in the charge $-q$ from infinity up to C without letting the charges at A and B move, then the work necessary will be $W_2 = \frac{1}{4\pi\epsilon_0} \left(\frac{q^2}{\sqrt{2}a} - \frac{q^2}{a} \right)$. Finally, the work necessary to bring in the charge q from infinity up to D in the presence of the other three charges will be $W_3 = \frac{1}{4\pi\epsilon_0} \left(\frac{-q^2}{a} + \frac{q^2}{\sqrt{2}a} - \frac{q^2}{a} \right)$. In other words, the total work required to assemble the charges is $W = W_1 + W_2 + W_3 = -\frac{(4-\sqrt{2})q^2}{4\pi\epsilon_0 a}$.

The fact that W is negative implies that a positive amount of work is to be performed so as to *disperse* the given assembly of charges to infinite distances from one another.

More generally, one may consider an assembly of charges, with charges q_i located at points P_i , where the potential at P_i ($i = 1, 2, \dots, N$) due to the other charges of the assembly is V_i . The electrostatic energy W of such a system, which is the same thing as the work required to assemble it starting from a configuration where the charges at infinite distances from one another, is then

given by the expression

$$W = \frac{1}{2} \sum_{i=1}^N q_i V_i. \quad (11-14)$$

This formula can be derived by making use of the linear relationship between the potential at any point in an electric field and the source charges responsible for the field (in this context, see also sec. 11.9.3). In the present instance of the charges placed at the four corners of a square, verify that the expression (11-14) indeed leads to the expression $W = -\frac{(4-\sqrt{2})q^2}{4\pi\epsilon_0 a}$ obtained above by a step-by-step consideration of the process of assembly of the given system.

11.4.7 Deriving the intensity from the potential

Knowing the electric field intensity at various different points in an electric field, one can work out the potential at any given point from eq. (11-7). I now address the reverse problem of working out the intensity from the potential $V(\mathbf{r})$.

Consider two points P and P' close to each other, with position vectors \mathbf{r} and $\mathbf{r} + \delta\mathbf{r}$ respectively. The difference of potentials at these two points is given by

$$V_{P'} - V_P = -\mathbf{E}(\mathbf{r}) \cdot \delta\mathbf{r}. \quad (11-15a)$$

This relation is a consequence of the basic fact that the potential difference between two points (P' and P in the present instance) is the work done against the electrical force in transferring a unit charge from one point (P) to the other (P').

For the sake of convenience, let us denote the unit vector along $\delta\mathbf{r}$ as \hat{s} and use the notation $|\delta\mathbf{r}| = \delta s$. This means that the point P' is situated at a distance δs from P along the direction of the unit vector \hat{s} . Then, writing δV for $V_{P'} - V_P$, one has

$$\delta V = -\mathbf{E} \cdot \hat{s} \delta s. \quad (11-15b)$$

In this equation $\mathbf{E} \cdot \hat{s}$ stands for the component of the electric field intensity along \hat{s} , which we write as E_s . Further, assuming the distance δs to be infinitesimal, the ratio $\frac{\delta V}{\delta s}$ represents the rate of change of V with distance along the unit vector \hat{s} , i.e., in other

words, the *space derivative* of the potential along the chosen direction \hat{s} . It is referred to as the *directional derivative* of V along \hat{s} at the point P, and is written as $\frac{\partial V}{\partial s}$. One then has

$$E_s = -\frac{\partial V}{\partial s}. \quad (11-16)$$

In other words, *the component of intensity along any given direction is the directional derivative of the potential along that direction.*

Here I have used the notation $\frac{\partial V}{\partial s}$ to denote the directional derivative, and not $\frac{dV}{ds}$ because it represents the *partial derivative* of V with distance s along \hat{s} . One can describe the locations of points in space close to the point P with the help of three independent co-ordinates, of which one can be chosen as s (the other two being, say, u and v). The partial derivative with respect to s then means the rate of change with respect to s *with u and v held constant*.

Equation (11-16) does not straightaway give us the vector \mathbf{E} at the point P, instead specifying its component along any chosen direction \hat{s} . Evidently, knowing the component along various different directions in space, one gets to know the vector \mathbf{E} completely. In reality, it is sufficient to know *three* components along, say, three mutually perpendicular directions. Thus, choosing any Cartesian co-ordinate system with co-ordinates x , y , and z , one gets the components of intensity along the unit vectors \hat{i} , \hat{j} , and \hat{k} respectively as

$$E_x = -\frac{\partial V}{\partial x}, \quad E_y = -\frac{\partial V}{\partial y}, \quad E_z = -\frac{\partial V}{\partial z}. \quad (11-17)$$

Here $\frac{\partial V}{\partial x}$ is obtained by working out the rate of variation of V with x , with y and z held constant, this rate of variation being evaluated *at* the point P under consideration. The other two partial derivatives, $\frac{\partial V}{\partial y}$ and $\frac{\partial V}{\partial z}$ can be evaluated in a similar manner.

In mathematical terms, one expresses equations (11-17) in the compact form

$$\mathbf{E} = -\nabla V, \quad (11-18)$$

where ∇V is termed the *gradient* of V (see section 2.14.1 for an introduction to the concept of the gradient of a scalar field). Thus we arrive at the result that *the electric intensity at any given point P is the gradient of the potential at that point, taken with a negative sign.*

In particular, for a *uniform* electric field, say along the x-axis, if V_1 and V_2 denote the potentials at two points separated along the x-axis by a distance d , then intensity along the x-axis is given by (see eq. (11-9), which expresses the same result)

$$E = -\frac{V_2 - V_1}{d}. \quad (11-19)$$

Problem 11-5

Imagine a circular distribution of charge where a charge q is distributed uniformly on one quarter of the circumference of a circle C of radius R , while a charge $-3q$ is distributed uniformly on the rest of the circumference. Calculate the potential at the centre O of the circle as also at a point P located at a distance z from O, on a line perpendicular to the plane of C and passing through O. Obtain the component of the electric field intensity at P along the line OP.

Answer to Problem 11-5

.

Problem: .

Hint: Since all points on the circle C are equidistant from O, as also from P, it does not really matter as to how the charge is distributed on C so far as the potential at O or P is concerned, and what matters is simply the total charge $Q = -2q$ (reason this out, making use of the principle of superposition). The potential is thus of the form $V = -\frac{2q}{4\pi\epsilon_0 d}$ where $d = R$ for the point O, and $d = \sqrt{(R^2 + z^2)}$ in the case of P. The component of electric field intensity along OP at P is given by $E = -\frac{dV(z)}{dz}$, where $V(z) = -\frac{2q}{4\pi\epsilon_0 \sqrt{(R^2 + z^2)}}$.

11.5 Force on a thin layer of charge

Fig. 11-7(A) depicts a small element A of a thin charged layer, while a part B of the surrounding portion of the layer is shown with dotted lines. This element of area experiences a force due to the electric field produced by the layer, which we now evaluate.

Let the surface charge density (i.e., the charge per unit area; refer to definition in formula (11-71) below where the ratio is considered in the limit of a vanishingly small area) and field intensity at an point on the layer be denoted by σ and \mathbf{E} , where these may vary from point to point. The expression for the force on the small portion A would then seem to be $\sigma \mathbf{E} \delta S$, where δS stands for its surface area. However, this expression needs a modification, since a part of the electric field strength \mathbf{E} is produced by the charge in A itself. In other words, in calculating the force on A, we have to discount its *own* field. The latter is oppositely directed on the two sides of A, and thus produces a *discontinuity* in the field \mathbf{E} , while the field strength caused by the *rest* of the layer involves no such discontinuity.

Put differently, the total field \mathbf{E} is made up of two parts

$$\mathbf{E} = \mathbf{E}' + \mathbf{E}'', \quad (11-20)$$

where the ‘self-field’ (\mathbf{E}') is directed oppositely at points Q and R on the two sides of the layer, located close to A (refer to fig. 11-7(B); in the figure, the charge in A is assumed to be positive for the sake of concreteness), while the field \mathbf{E}'' , caused by the rest of the layer has the same value at the two points (in the limit of vanishingly small area of the element A; the direction of \mathbf{E}'' shown in the figure is chosen arbitrarily).

One can thus write

$$\mathbf{E}'' = \frac{1}{2}(\mathbf{E}_Q + \mathbf{E}_R) = \mathbf{E}_{av}, \quad (11-21)$$

i.e., the field \mathbf{E}'' , which is to be used in the calculation of the force experienced by the element A of area δS , is nothing but the *average* field across the layer (reason this out;

refer to fig. 11-7(B); in the expressions for \mathbf{E}_Q and \mathbf{E}_R , the part \mathbf{E}'' remains the same, while \mathbf{E}' assumes equal and opposite values).

Thus, finally, the force experienced by the element A of area δS can be expressed as

$$\delta \mathbf{F} = \sigma \mathbf{E}_{\text{av}} \delta S, \quad (11-22)$$

i.e., the *force per unit area* (the surface density of force) acting on the thin layer of charge is given by

$$\mathbf{f} \equiv \frac{\delta \mathbf{F}}{\delta S} = \sigma \mathbf{E}_{\text{av}} = \frac{1}{2} \sigma (\mathbf{E}_- + \mathbf{E}_+). \quad (11-23)$$

Here and in fig. 11-7(B), the three terms appearing in eq. (11-20) pertaining to the two sides of the layer of charge are distinguished by suffixes '-' and '+' (thus, $\mathbf{E}'_- = -\mathbf{E}'_+$, $\mathbf{E}''_- = \mathbf{E}''_+$, $\mathbf{E}_- = \mathbf{E}_Q$, $\mathbf{E}_+ = \mathbf{E}_R$).

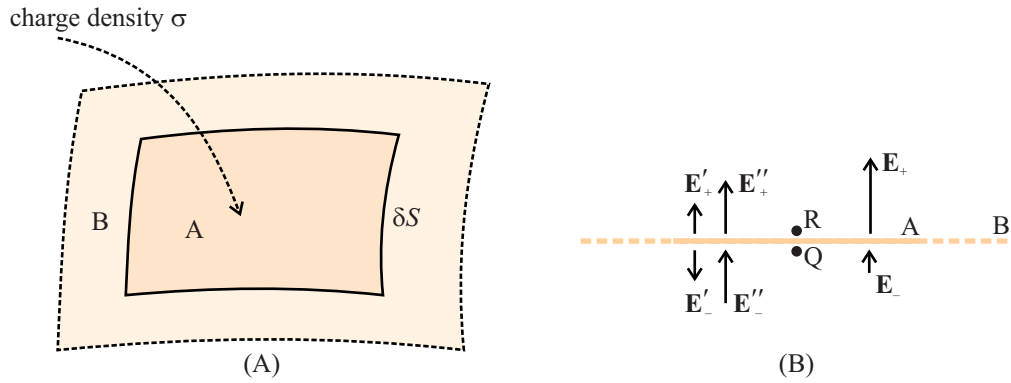


Figure 11-7: (A) A small element of area δS in a thin layer of charge, where the surface charge density is σ ; a surrounding part of the element A is shown with dotted lines; (B) a section by a plane perpendicular to the plane of the element A, shown for the sake of convenience of representation; Q and R are points on the two sides of the element A, located close to it, separated from each other by a vanishingly small distance; the self-field \mathbf{E}' at the two points are equal and opposite, these being denoted by \mathbf{E}'_- and \mathbf{E}'_+ ; the field \mathbf{E}'' , caused by the rest of the charged layer, is continuous across it, and hence, is the same at the two points; this is the same as the average total field $\mathbf{E}_{\text{av}} = \frac{1}{2}(\mathbf{E}_- + \mathbf{E}_+)$ across the layer; the force per unit area on A, caused by the field produced by the charged layer, is given by the expression (11-23).

We will find an application of this result in sec. 11.10.3.2 below.

11.6 Electric dipole and dipole moment

11.6.1 A pair of equal and opposite point charges

In fig. 11-8 below, two source charges, $-q$ and $+q$ ($q > 0$), are located at the points A and B, which we assume to be on the x-axis of a Cartesian co-ordinate system, at equal distances on either side of the origin O. We assume, moreover, that the field point P is located in the x-y plane of the co-ordinate system (given the two source charges and the field point, the co-ordinate system can always be chosen so as to conform to these requirements). If the distance between the source charges be d , then the co-ordinates of A and B in the x-y plane will be given by $(-\frac{d}{2}, 0)$ and $(\frac{d}{2}, 0)$ respectively. If the co-ordinates of the field-point be (x, y) then its distance from the two source points will be, respectively,

$$u_1 = \left(\left(x + \frac{d}{2} \right)^2 + y^2 \right)^{\frac{1}{2}}, \quad u_2 = \left(\left(x - \frac{d}{2} \right)^2 + y^2 \right)^{\frac{1}{2}}. \quad (11-24)$$

One can then determine the potential at P by making use of eq. (11-13b):

$$V = \frac{1}{4\pi\epsilon_0} \left(-\frac{q}{u_1} + \frac{q}{u_2} \right). \quad (11-25)$$

The distance of the field point P from the origin O is

$$r = (x^2 + y^2)^{\frac{1}{2}}, \quad (11-26a)$$

which we assume to be large compared to the distance d between the two source charges:

$$r \gg d. \quad (11-26b)$$

One can then invoke the binomial expansion in (11-24) to obtain

$$u_1 \approx r + \frac{xd}{2r}, \quad u_2 \approx r - \frac{xd}{2r}, \quad (11-26c)$$

where the \approx symbol has been used to denote an approximate equality. According to the rules of binomial expansion, the right hand sides of equations (11-26c) should contain

additional terms, but all those have been ignored on ground of (11-26b) since these are of the order of $(\frac{d}{r})^2$ or even smaller.

If $\frac{d}{r}$ is a small quantity, then $(\frac{d}{r})^2$, $(\frac{d}{r})^3$, etc., are of progressively higher orders of smallness. For instance, if $\frac{d}{r}$ is of the order of 10^{-4} , then $(\frac{d}{r})^2$ is of the order of 10^{-8} , and $(\frac{d}{r})^3$ is of the order of 10^{-12} . Terms involving $(\frac{d}{r})$, $(\frac{d}{r})^2$, ... are said to be of the first, second, ... orders of smallness respectively. Equations (11-26c) are thus obtained by retaining terms of the first order of smallness while ignoring terms of second and higher orders. More precisely, even if the second order terms were retained, the result (11-27) below would remain valid, since these would get canceled in the end. In other words, the result we are going to derive is correct up to and including the second order terms. Evidently, the smaller the value of $\frac{d}{r}$ is, the better will the approximation be.

Substituting in eq. (11-25) and once again ignoring terms of second order of smallness, one obtains

$$\begin{aligned} V &\approx \frac{1}{4\pi\epsilon_0} \frac{(qd)x}{r^3} \\ &= \frac{1}{4\pi\epsilon_0} \frac{p \cos \theta}{r^2}. \end{aligned} \quad (11-27)$$

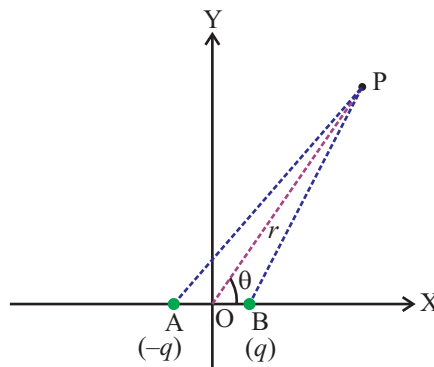


Figure 11-8: Potential due to a dipole, made up of point charges $-q$ and q located at the points A and B; P represents the field point; co-ordinate axes are chosen as shown.

Here, in the last expression, p stands for the product qd , and θ is the angle between OP and the line joining the two points A and B (the x-axis in the figure).

Notice that the total charge of the system made up of the two source charges $-q$ and $+q$ is zero, and the contributions of the two to the potential are of opposite signs. The magnitudes of the two contributions are close to each other (see eq. (11-25)) since, for $d \ll r$, u_1 and u_2 are nearly equal. The smaller the value of $\frac{d}{r}$, the more accurately do the two contributions to V cancel each other, leaving out a relatively small residual contribution given by eq. (11-27), where this residual contribution arises due to the fact that u_1 and u_2 are not exactly equal to each other.

11.6.2 Dipole moment and dipole

Denoting the unit vector along the line AB joining the two source charges by \hat{n} (which is \hat{i} in the situation depicted in fig. 11-8), the quantity $p\hat{n}$ is referred to as the *dipole moment* of the system of charges under consideration (at times the term dipole moment is used to denote the magnitude p). The *axis* of the dipole is said to be along the direction of the vector \hat{n} .

Denoting the dipole moment by \mathbf{p} , and the position vector of the field point P with respect to the origin O (the location of the mid-point of the two source charges) by \mathbf{r} , eq (11-27) can be written in the alternative form

$$V = \frac{1}{4\pi\epsilon_0} \frac{\mathbf{p} \cdot \mathbf{r}}{r^3}. \quad (11-28)$$

Observe that the dipole moment is a vector quantity and that its unit in the SI system is C·m. More generally, if \mathbf{r}_1 and \mathbf{r}_2 be the position vectors of the two source charges $-q$ and q respectively with respect to any chosen origin, then one has

$$d\hat{n} = \mathbf{r}_2 - \mathbf{r}_1, \quad (11-29a)$$

and

$$\mathbf{p} = (-q)\mathbf{r}_1 + q\mathbf{r}_2. \quad (11-29b)$$

The potential at any chosen field point will continue to be given by the formula (11-28), provided that \mathbf{r} is interpreted as the position vector of the field point with respect to the mid-point of the locations of the two charges. It may be mentioned here that a dipole moment can be defined for *any* system of charges, not necessarily made up of a pair of equal and opposite charges like the one being considered in the present section. Equation (11-29b) can then be seen as just a particular instance of the general expression of dipole moment for a system of charges, which I will state below (see eq. (11-31)).

A system made up of a pair of equal and opposite charges such as the one considered above is termed an *electric dipole*. The molecules of certain materials can be looked upon as dipoles in this sense, and a number of characteristic properties of these substances derive from the dipole moments of their molecules.

Strictly speaking, however, the definition of an electric dipole differs from above. Recall that eq. (11-27) is only an approximate and not a strict equality because it is obtained by ignoring terms of higher orders of smallness in the parameter $\frac{d}{r}$. As I have already mentioned, the error introduced by dropping these terms becomes progressively smaller as the ratio $\frac{d}{r}$ is made smaller. If now one imagines d to go to zero, and at the same time the vector \mathbf{p} to have a definite magnitude and direction (this necessarily requires that one has to simultaneously go to the limit $|q| \rightarrow \infty$) then eq. (11-27) will no longer require any correction and will be an exact one. *A pair of charges imagined to satisfy these requirements is said to constitute an electric dipole.*

In other words, if two charges $-q$ and q are placed at a distance d from each other as in fig. 11-8, and if the limits $d \rightarrow 0$, $|q| \rightarrow \infty$, and $qd \rightarrow p$ are satisfied with \hat{n} remaining a fixed unit vector, then these two make up an electric dipole, while $\mathbf{p} = p\hat{n}$ is termed its dipole moment. At times, the term dipole moment is used to refer to the magnitude $|p|$ of \mathbf{p} .

However, as I have mentioned above, the term electric dipole is often used to refer to a pair of equal and opposite charges where the above limits are not necessarily satisfied (in practice, though, the term is used when the distance d is sufficiently small, as in the case of a molecule). In that case the dipole moment of a pair of charges $-q$ and q is defined as (see eq. (11-29b))

$$\mathbf{p} = q(\mathbf{r}_2 - \mathbf{r}_1), \quad (11-30)$$

where \mathbf{r}_1 and \mathbf{r}_2 denote the position vectors of the two charges respectively. At times, a pair of charges satisfying the limiting condition mentioned above is referred to as an *ideal* dipole so as to distinguish it from a *real* dipole with the charges separated by a small but finite distance.

In this context, one has to note that, while the term 'dipole' is employed to denote a pair of equal and opposite charges, the term 'dipole moment' applies to *any* system of charges whatsoever.

For instance, consider a system made up of point charges q_1, q_2, \dots, q_N , located at points with position vectors $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N$ with respect to any chosen origin. then the vector

$$\mathbf{p} = q_1\mathbf{r}_1 + q_2\mathbf{r}_2 + \dots + q_N\mathbf{r}_N = \sum_{i=1}^N q_i\mathbf{r}_i, \quad (11-31)$$

is referred to as the *dipole moment* of the system. Evidently, eq. (11-30) is a particular instance of this definition, when the system under consideration comprises of a pair of equal and opposite charges.

Potential due to a system of charges. Significance of dipole moment.

We know that the potential and intensity due to the above system of charges at a field point with position vector \mathbf{r} are given by expressions (11-13b) and (11-6a) respectively. Under certain circumstances, however, these expressions can be simplified in terms

of the total charge Q of the system

$$Q = \sum_{i=1}^N q_i, \quad (11-32)$$

and its dipole moment \mathbf{p} as defined in eq. (11-31), where certain *approximations* are to be made in arriving at these simplifications.

For instance, suppose that the system of charges under consideration is contained within a finite region of space (R) of linear dimension d , as in fig. 11-9. In the present context, the quantity d need not be defined exactly. Considering the pairwise distance between the charges constituting the system, d may be defined to be the largest of these distances. It may also be taken to be slightly larger or smaller than this maximum separation without materially altering the conclusions presented below. In an approximation scheme, the orders of magnitude of a number of quantities are found to be more relevant than their exact values.

For the sake of concreteness, we assume that the origin O is located within the region R. Once again, the exact location of the origin does not really matter in our approximation scheme. We consider the potential and intensity due to the above system of source charges at the field point P with position vector, say, \mathbf{r} with respect to the origin, where $|\mathbf{r}|$ is large compared to d :

$$d \ll |\mathbf{r}|. \quad (11-33)$$

The potential at such a large distance is then given by the expression

$$V(\mathbf{r}) \approx \frac{1}{4\pi\epsilon_0} \left(\frac{Q}{r} + \frac{\mathbf{p} \cdot \mathbf{r}}{r^3} \right), \quad (11-34)$$

where Q stands for the total charge of the system given by eq. (11-32) and \mathbf{p} is the dipole moment given by (11-31).

Note that (11-34) is not an exact equality. A number of correction terms are to be added to its right hand side so as to arrive at an equality. However, all these correction

terms are small, and can be ignored, if (11-33) is satisfied, i.e., the field point is located at a sufficiently large distance from the system of charges under consideration.

Eq. (11-34) tells us that, under condition (11-33), the potential due to the system of charges under consideration is effectively the superposition of the potential due to a single point charge Q and that due to an ideal dipole of dipole moment \mathbf{p} placed at the origin.

Knowing the potential from eq. (11-34) one can determine the field intensity \mathbf{E} by working out the gradient of the potential, where the Cartesian components of the intensity are given by equations (11-17). The expression for the intensity will once again correspond a superposition of two terms : the intensity due to a point charge, given by an expression of the form (11-5a), and that due to an ideal dipole. This latter turns out to be given by the expression (refer to problem 11-6)

$$\mathbf{E}_{\text{dipole}}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \left(\frac{3(\mathbf{p} \cdot \mathbf{r})\mathbf{r} - r^2\mathbf{p}}{r^5} \right). \quad (11-35)$$

The approximate expression for the intensity due to a system of localized charges at a large distance is then found to be

$$\mathbf{E} \approx \frac{1}{4\pi\epsilon_0} \left(\frac{Q}{r^3} \mathbf{r} + \frac{3(\mathbf{p} \cdot \mathbf{r})\mathbf{r} - r^2\mathbf{p}}{r^5} \right). \quad (11-36)$$

Having obtained the approximate expressions for the potential and intensity due to a system of localized charges at a distant field point, let us take note of the following relevant observations.

1. If the total (or net) charge Q of the system under consideration is not zero, i.e., the positive and negative charges in it do not cancel one another, the second term in the right hand side of eq. 11-34 (or of (11-36)) becomes smaller than the first term by a factor of the order of $\frac{d}{r}$ (check this out) and hence can be ignored. In other words, if the net charge (also referred to as the *monopole moment* of the system under consideration) is non-zero, then the system can effectively be replaced

with a single point charge Q so far as the potential and intensity at distant field points are concerned. Evidently, this constitutes a great simplification in the determination of the potential and intensity due to a system of charges.

2. If the net charge, or monopole moment Q of the system happens to be zero, i.e., the positive and negative charges in it cancel one another, then the first term in the right hand side of eq. (11-34) (or of (11-36)) becomes zero, and consequently, the approximate expression for the potential or intensity *reduces to that due a single dipole placed at the origin*. Once again, this is a convenient simplification for the potential and intensity at a distant field point for a system with zero monopole moment.

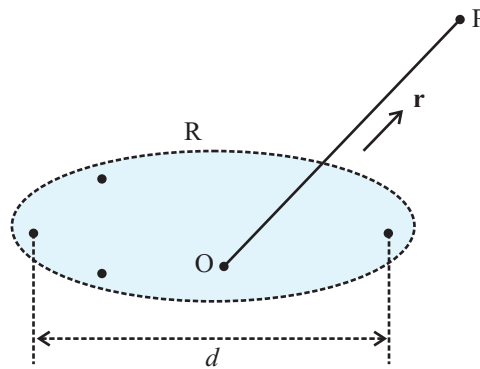


Figure 11-9: Potential due to a system of localized charges; the source charges are all located within a region R of dimension d ; the field point P is at a large distance r from the origin O; the potential at P is then determined, in an approximate sense, solely by the total charge and the dipole moment of the system of source charges.

Problem 11-6

Establish the validity of formula 11-35.

Answer to Problem 11-6

Refer to fig. (11-8). The x- and y-components of intensity at P due to the pair of charges is obtained by making use of eq. (11-6a), each as a sum of two terms:

$$E_x = \frac{q}{4\pi\epsilon_0} \left(\frac{x - \frac{d}{2}}{u_2^3} - \frac{x + \frac{d}{2}}{u_1^3} \right), \quad E_y = \frac{q}{4\pi\epsilon_0} \left(\frac{y}{u_2^3} - \frac{y}{u_1^3} \right), \quad (11-37)$$

where u_1 and u_2 are defined in (11-26c). Employing the rules for binomial expansion and ignoring terms of the order of $(\frac{d}{r})^2$ and those of a higher degree of smallness, one obtains

$$E_x = \frac{1}{4\pi\epsilon_0} \left(\frac{p(3x^2 - r^2)}{r^5} \right), \quad E_y = \frac{1}{4\pi\epsilon_0} \left(\frac{3pxy}{r^5} \right). \quad (11-38)$$

Finally, check that this is the same as eq. (11-35) for the special choice of the co-ordinate system in fig. 11-8. Note that the expression (11-35) is an approximate one, valid under the condition (11-26b).

Problem 11-7

Charges $-q$, q , $-q$, q are placed at points $(-a, 0)$, $(a, 0)$, $(0, -a)$, $(0, a)$ respectively, in a plane with reference to a Cartesian co-ordinate system, where $q = 4.8 \times 10^{-16} \text{C}$, and $a = 1.0 \times 10^{-8} \text{m}$. Estimate the potential due to this system of charges at the point $(\frac{b}{2}, \frac{\sqrt{3}}{2}b)$, where $b = 10^{-3} \text{m}$.

Answer to Problem 11-7

A good estimate of the potential at the given point (P, say) is obtained by noting that the system of charges can be replaced with two electrical dipoles, each of moment $p = 9.6 \times 10^{-24} \text{C}\cdot\text{m}$, placed at the origin, where one is oriented along the unit vector \hat{i} and the other along \hat{j} (reason this out). Making use of the formula (11-27) (or, equivalently, (11-28)) along with the principle of superposition, the potential at P is seen to be

$$V = \frac{p}{4\pi\epsilon_0 d^2} (\cos \theta_1 + \cos \theta_2),$$

where $d = b$, and $\theta_1 = \frac{\pi}{3}$, $\theta_2 = \frac{\pi}{6}$, the latter two being the angles made by the vector $\frac{b}{2}\hat{i} + \frac{\sqrt{3}b}{2}\hat{j}$ with \hat{i} and \hat{j} respectively. Substituting given values, we get $V = 1.18 \times 10^{-7} \text{V}$.

11.6.3 Torque and force on a dipole in an electric field

11.6.3.1 Torque on a dipole

Fig. 11-10 depicts a pair of equal and opposite charges ($-q$ and q) placed at points P and Q respectively, in a uniform electric field of strength \mathbf{E} , where the separation between P and Q is d , the unit vector from P to Q being, say, \hat{n} .

The forces on the two charges due to the field are, respectively, $-q\mathbf{E}$ and $q\mathbf{E}$. Since these two forces are equal and opposite to each other, the net force on the pair of charges is zero, and hence, the sum of the moments of the forces about any point will be independent of the location of that point (refer to section 3.22.3.3). Indeed, the two forces constitute a couple, where the moment of the couple is the sum of the two moments mentioned above. Taking the moments about the point P, one obtains the moment of the couple to be

$$\mathbf{M} = (qd)\hat{n} \times \mathbf{E}. \quad (11-39)$$

The system of two charges under consideration may be looked upon as a real (as distinct from an ideal) dipole, with dipole moment $\mathbf{p} = (qd)\hat{n}$, and the above result then tells us that the moment of the couple acting on the dipole, or the *torque* on the dipole placed in a uniform electric field is given by

$$\mathbf{M} = \mathbf{p} \times \mathbf{E}. \quad (11-40)$$

Suppose now that the field is a *non-uniform* one while, at the same time, imagine the point Q to be made to tend to P (i.e., $d \rightarrow 0$) and q to be made correspondingly large so that, in the limit, the two charges under consideration make up an *ideal* dipole of dipole moment $\mathbf{p} = \hat{n} \lim (qd)$ placed at P. The resultant moment of the electrical forces on the system is then again given by eq. (11-39) with the above limit understood, and with \mathbf{E} now standing for the electric field intensity *at* P.

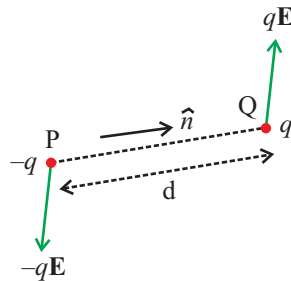


Figure 11-10: Dipole in uniform electric field - illustrating torque on a dipole.

In other words, the torque on an ideal dipole of moment \mathbf{p} placed in an electric field that need *not* be a uniform one is given by the expression (11-40), where now \mathbf{E} stands for the electric field intensity at the location of the dipole.

Thus, the magnitude of the torque is $pE \sin \theta$, where θ is the angle between the dipole axis (the directed line along \hat{n}), and the direction of the electric field intensity. Evidently, then, the torque on the dipole will be zero when the dipole is aligned either parallel or anti-parallel to the field.

11.6.3.2 Force on a dipole

As seen above, the net force on a dipole (whether a real or an ideal one) placed in a uniform electric field is zero, since the forces on the two constituent charges are equal and opposite. For a *non-uniform* electric field, on the other hand, the two forces do not cancel each other, and there acts a resultant force on the dipole.

Considering, for instance, an ideal dipole of moment p oriented along the x-axis, the force on the dipole works out to

$$\mathbf{F} = p \left(\hat{i} \frac{\partial E_x}{\partial x} + \hat{j} \frac{\partial E_y}{\partial x} + \hat{k} \frac{\partial E_z}{\partial x} \right). \quad (11-41)$$

In this expression, $\frac{\partial E_x}{\partial x}$, $\frac{\partial E_y}{\partial x}$, and $\frac{\partial E_z}{\partial x}$ are the *partial derivatives* of the three Cartesian components of \mathbf{E} with respect to x , *evaluated at the location of the dipole* where, in evaluating these derivatives, y and z are assumed to be held constant.

In summary, a dipole placed in an electric field experiences a force \mathbf{F} given by eq. (11-41) (where the x-axis has been chosen along the direction of the dipole moment; the line of action of the force passes through the point of location of the dipole) and a couple of moment \mathbf{M} given by eq. (11-40).

The force on the dipole can be expressed in the more general form

$$\mathbf{F} = \hat{i} \frac{\partial(\mathbf{p} \cdot \mathbf{E})}{\partial x} + \hat{j} \frac{\partial(\mathbf{p} \cdot \mathbf{E})}{\partial y} + \hat{k} \frac{\partial(\mathbf{p} \cdot \mathbf{E})}{\partial z}. \quad (11-42)$$

This expression is seen to reduce to $(p_x \frac{\partial}{\partial x} + p_y \frac{\partial}{\partial y} + p_z \frac{\partial}{\partial z})(\hat{i}E_x + \hat{j}E_y + \hat{k}E_z)$ when one makes use of the fact that the electric field is a conservative one, i.e., the electric field intensity can be expressed in terms of a scalar potential. In particular, one obtains the expression in (11-41) when $\mathbf{p} = \hat{i}p$.

11.6.4 Potential energy of a dipole in an electric field

Consider a dipole in an electric field where it is assumed that the dipole can be made to undergo a translational motion from one point to another and also a rotation about any given axis. One can imagine the dipole to be a pair of charged particles joined rigidly to each other such that the magnitude of the dipole moment (p) is fixed, but its position and orientation can vary under the influence of the field. At any given instant, the position and orientation of the dipole can be completely specified in terms of five independent variables, namely, three position co-ordinates (x, y, z) with respect to any given Cartesian co-ordinate system, and two angles (say, θ, ϕ) defining the orientation of the dipole.

Referring to the instantaneous motion of the dipole at any given time t , these two angles can be chosen as in fig. 11-11, where the z -axis of the above co-ordinate system is chosen to be along the electric field intensity at the location of the dipole, and θ is the angle made by the dipole with the z -axis. The remaining angle ϕ can then be chosen as the angle between the z - x plane and the plane containing the dipole axis and the z -axis (ϕ is then referred to as the *azimuthal angle* in a *spherical polar co-ordinate system*).

Thus, any infinitesimal translation and rotation of the dipole in the electric field is described by the five independent quantities $\delta x, \delta y, \delta z, \delta \theta$, and $\delta \phi$, of which the former three give the translations along the three co-ordinate axes. The work done in such a translation and rotation by the force and torque acting on the dipole can be expressed

in the form

$$\delta W = F_x \delta x + F_y \delta y + F_z \delta z - M_x \sin \phi \delta \theta + M_y \cos \phi \delta \theta + M_z \delta \phi. \quad (11-43)$$

Here the first three terms represent the work done in the translational motion of the dipole while the last three terms represent the work done in its rotational motion (for which, refer to the expression (3-167) where the notation differs slightly; the infinitesimal angles of rotation about the three axes for given infinitesimal changes in θ , ϕ turn out to be, respectively, $-\sin \phi \delta \theta$, $\cos \phi \delta \theta$, and $\delta \phi$). It may be mentioned that the expression (11-40) shows that M_z is zero (reason this out), i.e., the last term in eq. (11-43) is, in reality, redundant.

Considering any finite motion of the dipole from any given initial to a final configuration through a succession of intermediate configurations (each configuration being completely specified by the values of the five variables mentioned above) the work done on it by the field can be expressed by summing up expressions of the form (11-43) over the successive infinitesimal translations and rotations through which the final configuration is reached.

The electric field being *conservative* in nature, this work has to be independent of the series of intermediate configurations through which the dipole passes, being determined solely by the initial and final configurations of the dipole.

The mathematical criterion necessary for such a requirement to hold good is that there must exist a function U , the *potential energy* of the dipole in the electric field, such that the components of the forces and moments can be expressed in terms of its *partial*

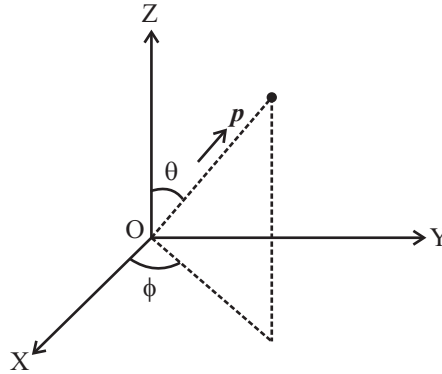


Figure 11-11: Dipole of moment \mathbf{p} in an electric field; the dipole axis (along \mathbf{p}) makes an angle θ with the z -axis, chosen along the direction of the field, while the azimuthal angle is ϕ ; one can choose the co-ordinate axes such that the dipole axis lies in the $z - x$ plane at any given time instant, in which case $\phi = 0$ (the figure, however, shows a more general orientation with a different value of ϕ); the instantaneous rotational motion of the dipole then consists of a rotation about the z -axis and another about the y -axis.

derivatives, all evaluated at the instantaneous configuration of the dipole:

$$\begin{aligned} F_x &= -\frac{\partial U}{\partial x}, \quad F_y = -\frac{\partial U}{\partial y}, \quad F_z = -\frac{\partial U}{\partial z}, \\ -M_x \sin \phi + M_y \cos \phi &= -\frac{\partial U}{\partial \theta}, \quad M_z = -\frac{\partial U}{\partial \phi}. \end{aligned} \quad (11-44)$$

A few steps of mathematical reasoning tells us (however, I am going to skip the derivation here) that this function U indeed exists, and is given by

$$U = -\mathbf{p} \cdot \mathbf{E} = -pE \cos \theta. \quad (11-45)$$

This expression gives the potential energy of the dipole in an electric field, where θ stands for the angle between the dipole axis and the direction of the electric field at the location of the dipole (note from eq. (11-44) and (11-45) that $M_z = 0$, as required by (11-40)). The work done on the dipole (expression (11-43)) by the electrical forces then relates to the change in its potential energy as

$$\delta W = -\delta U. \quad (11-46)$$

Note that the formulae (11-44), (11-45) are consistent with the expressions (11-40),

(11-42) obtained earlier.

Problem 11-8

Consider a point charge $q = 1.0 \times 10^{-16} \text{C}$ fixed at the origin (O) of a Cartesian co-ordinate system and a dipole of moment $p = 5.0 \times 10^{-19} \text{C}\cdot\text{m}$ placed at a point P with co-ordinates $(D = 2.0 \times 10^{-8}, 0, 0) \text{m}$, pointing along a direction parallel to the z-axis. Find the potential energy of the dipole, and the force and torque experienced by it.

Answer to Problem 11-8

The electric field intensity at any point A with co-ordinates (x, y, z) due to the point charge q at O is given by $\mathbf{E} = \frac{q}{4\pi\epsilon_0 r^3}(x\hat{i} + y\hat{j} + z\hat{k})$, where $r = \sqrt{x^2 + y^2 + z^2}$ (draw a figure, if necessary). The force on the dipole of moment $\mathbf{p} = p\hat{k}$ placed at A would be, by formula (11-42), $\mathbf{F} = C(\hat{i}\frac{\partial}{\partial x}(\frac{z}{r^3}) + \hat{j}\frac{\partial}{\partial y}(\frac{z}{r^3}) + \hat{k}\frac{\partial}{\partial z}(\frac{z}{r^3}))$, where $C = \frac{qp}{4\pi\epsilon_0}$. On evaluating the derivatives and taking the point A to be at P (i.e., putting $y = z = 0, x = r = D$), one obtains $\mathbf{F} = \frac{C}{D^3}\hat{k} = \frac{qp}{4\pi\epsilon_0 D^3}\hat{k}$ (the same result is obtained from formula(11-41), but with the partial derivatives $\frac{\partial}{\partial x}$ replaced with $\frac{\partial}{\partial z}$ since, in writing (11-41), the dipole was assumed to be oriented along the x-axis). Using given numerical values, the force, acting along the z-axis, works out to $F = 0.056 \text{N}$. The torque is obtained from formula (11-40) by taking $\mathbf{p} = p\hat{k}$, and $\mathbf{E} = \frac{q}{4\pi\epsilon_0} \frac{1}{D^2}\hat{i}$, the latter being the field strength at P due to the charge q placed at O. This works out to a torque about the y-axis, of magnitude $1.12 \times 10^{-9} \text{N}\cdot\text{m}$. The potential energy of the dipole is, by formula (11-45), zero, since the dipole axis is oriented at an angle $\theta = \frac{\pi}{2}$ to the direction of the field strength. the torque about the y-axis tends to align the dipole with the direction of the field intensity (i.e., along the x-axis), thereby reducing its potential energy.

11.7 Electric lines of force and equipotential surfaces

11.7.1 Geometrical description of an electric field. Neutral points.

The electric field intensity (also referred to as the field intensity, see sec. 11.10.5.3) at any point in an electric field, being a vector quantity, is represented by a directed line segment. One can then think of the set of all such directed line segments at the various different points in the field. Taken together, these specify a *vector field* in the sense

mentioned in section 2.13, which can be taken to constitute a geometrical description of the electric field. A related geometrical description makes use of the concept of *electric lines of force*.

In fig. 11-12, a curve ABC has been drawn in an electric field in accordance with the following rule: the electric field intensity \mathbf{E} at any point P on the curve is directed along the tangent to it drawn at P. If the intensity at each point on the curve ABC is similarly directed, then this is referred to as a *line of force* in the electric field.

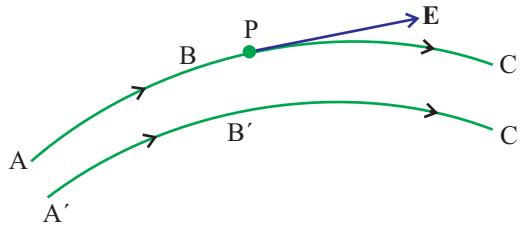


Figure 11-12: Illustrating the concept of lines of force in an electric field: two lines of force (ABC and A'B'C') are shown; the tangent to any line of force at any given point on it (such as the point P on ABC) gives the direction of the electric field intensity at that point; each line of force can be assigned a direction to denote the general direction in which the electric field intensity vector points as one moves along this line.

More generally, such lines can be drawn so as to provide a geometrical description of any vector field, the lines of force of an electric field being a particular instance of such a description.

In the figure, A'B'C' is another such line of force in the electric field, drawn beside ABC. The intensity at any point of this line is again directed along the tangent to A'B'C' drawn at that point. One can, in general, draw a line of force through each point in an electric field, and the set of lines of force so obtained constitutes a geometrical description of the field.

However, if the intensity at any point in the field be zero then no line of force can pass through that point. Such a point in an electric field is referred to as a *neutral point*. Since the intensity at a neutral point is zero, no specific direction can be associated

with such a point, which thereby appears as an exceptional or *singular* point in the field, looked upon as a vector field. Singular points other than neutral points also exist, such as the locations of point-like source charges.

Two instances of the disposition of lines of force in an electric field are shown schematically in fig. 11-13(A) and (B). In fig. 11-13(A), the lines of force in the vicinity of a point charge at the point O are shown, it being assumed that no other source charge affects the field in this vicinity. All the lines of force here are straight lines radiating from the source charge since, according to eq. (11-5a) the field intensity due to a source charge is directed radially with the source charge at the center. Evidently, there can be no neutral points in the field of a single point charge, since for any arbitrarily chosen position of the field point P, there exists a field line passing through it along the radial direction. Notice that *at* the point O, neither the magnitude nor the direction of the field intensity is defined. It is a singular point of the field, but not a neutral point.

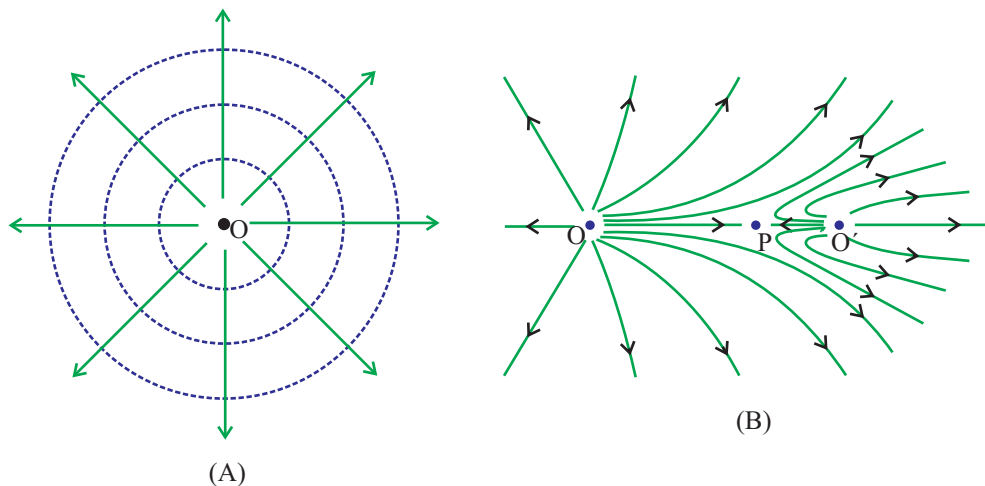


Figure 11-13: Lines of force in an electric field: (A) lines of force due to a single point charge at O; the lines of force are arranged radially with O at center; O is a singular point of the field; the dashed circles are sections of spherical equipotential surfaces to which the lines of force are directed normally; (B) lines of force due to two positive charges at O and O'; P is the neutral point, while O and O' are two singular points where the dispositions of the lines of force differ from the disposition near P.

Now look at fig. 11-13(B) where the field lines due to two positive charges q and q' ,

located at O and O' respectively, are shown. Assuming that there are no other source charges nearby, a point P on the line OO' will be a neutral point, this being the point dividing the segment OO' internally in the ratio $\sqrt{\frac{q}{q'}}$ (check this statement out).

Note that, in fig. 11-13B, the two source points O and O' are singular points of the field, but are not neutral points. Observe the different manners the field lines are arranged near the neutral point P and a singular point like O or O'.

11.7.2 Characteristics of lines of force

Two distinct lines of force in an electric field cannot cross each other. Looking at fig. 11-14 where two lines of force are imagined to cross each other at P, one observes that the electric intensity at P has to have two different directions, which is a contradiction in itself.

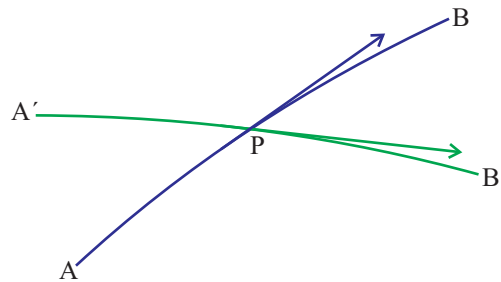


Figure 11-14: Depicting an imagined situation where two lines of force APB and A'PB' cross each other at P; such a crossing is impossible since there would then be two distinct directions for the intensity at P.

In this context, the source points and the neutral points in an electric field, which are exceptional or singular points in the field, deserve special mention. This can be explained with reference to fig. 11-13(A) and (B). In 11-13(A), all the lines of force are seen to emanate from the source point at O, but the point O itself is not located on any line of force. Each point in the vicinity of O, however, is located on a line of force. It is in this sense that a source point like O is an exceptional point because the intensity at O cannot be defined. If, on the other hand there exists a *continuous distribution* of charge instead of one or more point-like source charges in some region of space then

that region need not contain an exceptional point.

The source charge at O in fig. 11-13(A) has been assumed to be a positive one. In this case, all the lines of force are seen to radiate away from O in the form of a divergent bundle. In the case of a negative source charge, on the other hand, the lines of force converge on to the source point in the form of a convergent bundle. In general, the lines of force originate from positive source charges and terminate on negative source charges, and they are directed from points at higher potentials toward those at lower potentials. This last statement derives from the fact the potential difference between any two points, say, P and Q is the line integral of the electrical intensity from Q to P, taken with a negative sign - a consequence of eq. (11-7).

However, the initial and terminal points of a line of force may not necessarily be a positive or negative point charge. For instance, imagining a single positive point charge, with no other charge around, the lines of force emanating from the source move on to infinite distances and terminate at 'points at infinity'. In other words, either of the terminal points of a line of force may be at infinity.

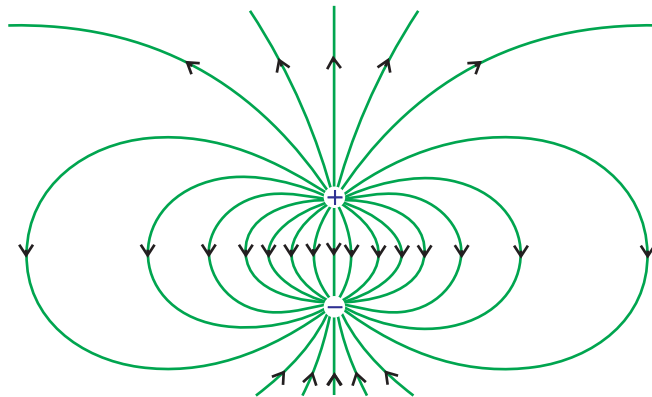


Figure 11-15: Lines of force due to an electric dipole; the two charges making up the dipole are shown separated from each other for the sake of illustration while, in reality, the dipole is a singular point source distinct from a monopole; the lines of force originate from and terminate at the location of the dipole; there are a pair of exceptional lines of force one of which originates from while the other terminates at infinity.

As I have already mentioned, the neutral points in an electric fields are also exceptional

points. In contrast to a source point, the intensity at such a point is well defined, namely, zero. This, however, means that the direction of the intensity at a neutral point is undefined. Once again, each point in a neighbourhood of a neutral point is located on some line of force or other, though no line of force passes through the neutral point itself. For instance, the neutral point P in fig. 11-13(B), located on the segment OO', does not have a line of force passing through it. The line of force from O to P and the one from P to O do not join continuously at O.

Fig. 11-15 depicts schematically a number of lines of force due to a dipole.

The expression for the intensity due to a dipole has been given in eq. (11-35). Recall that an ideal dipole is made up of a singular configuration of source charges, namely, a pair of positive and negative charges of infinite magnitude superposed on each other.

In other words, it is a point source, though not a point monopole.

Note in this figure that most of the lines of force originate *and* terminate at the point where the dipole is located whereas there are two exceptional lines of force, one originating from the dipole and terminating at infinity, and the other originating at infinity and terminating on the dipole.

11.7.3 Equipotential surfaces

An alternative geometrical description of an electric field is provided by its *equipotential surfaces*. Fig. 11-16 depicts schematically a set of equipotential surfaces in an electric field. Each of these surfaces is characterized by a fixed value of the potential at all points on it. In other words, all points on an equipotential surface are at the same potential. The figure shows three such surfaces with potentials V_1 , V_2 , and V_3 . Notice that the equipotential surfaces are *closed* ones. This is a general feature of the equipotential surfaces in an electric field, though there may exist one or more exceptional surfaces that extend to infinity. Each point in the field belongs to some equipotential surface or other. Thus, in fig. 11-16, there exist an infinite number of equipotential surfaces in between the ones with potentials V_1 and V_2 , and again between the ones with potentials

V_2 and V_3 .

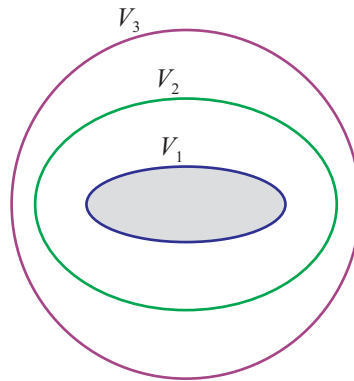


Figure 11-16: Illustrating the idea of equipotential surfaces; three surfaces corresponding to potentials V_1 , V_2 , and V_3 are shown, each of these being a closed surface; the potential is the same at all points on any equipotential surface.

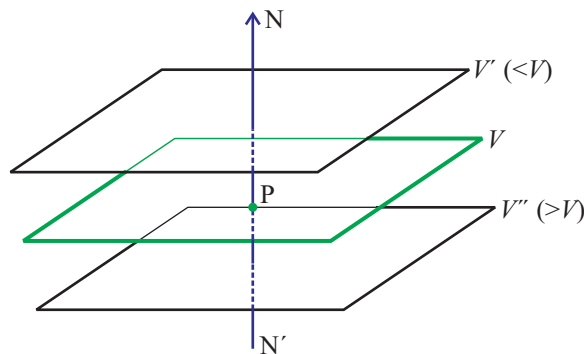


Figure 11-17: An equipotential surface through a point P in an electric field; only a small part of the surface is shown; the line of force through P lies along the normal NPN' to the equipotential surface through P, directed from P to N; parts of two other surfaces are also shown.

Recall that electric intensity and potential are both useful quantities in the mathematical description of an electric field. Correspondingly, the lines of force and equipotential surfaces are two convenient and complementary means for the geometrical description of the field. Analogous to the fact that two distinct lines of force cannot intersect, a pair of equipotential surfaces also cannot intersect each other. Consequently, the equipotential surfaces are, in general a family of non-intersecting closed surfaces. However, for a given value of the potential, say, V , one may have more than one closed surfaces

isolated from one another characterized by that value of the potential.

In fig. 11-17, P is a point in an electric field, through which passes an equipotential surface with potential V , only a part of the surface being shown in the figure. The normal to this surface at the point P is NPN' . Parts of equipotential surfaces with potentials $V'(< V)$ and $V''(> V)$ are also shown. In this case the tangent to the line of force passing through P will be along the normal NPN' and the line of force will be directed from P to N.

In other words, the relation between the equipotential surfaces and lines of force can be stated as follows: *all the lines of force are normal to the equipotential surfaces and are directed from higher to lower potentials.*

This geometrical relation of complementarity between the two is a reflection of the fact that the electrical intensity in a field is the gradient of the potential, taken with a negative sign.

As a simple instance of the above relation between the equipotential surfaces and lines of force, look at fig. 11-13(A), where the set of radial lines denote the lines of force due to a point charge at O. At the same time, the dashed spheres (circles in the plane of the figure) centred at O represent the equipotential surfaces, since all points at a given distance from a point charge correspond to the same value of the potential (refer to equations (11-11a) and (11-11b)). Evidently, the radial lines of force are all directed normally to the spherical equipotential surfaces.

11.7.4 Density of lines of force. Tubes of force.

Referring once again to fig. 11-13(A), let us consider a fixed number, say N , of lines of force emanating from the source point (the number shown in the figure is $N = 8$). Consider now two spherical surfaces centred around the source point, with radii, say, r_1 and r_2 ($r_1 < r_2$); such spherical surfaces are actually equipotential surfaces in the field, a number of which are shown with dashed circles in the figure). The surface area of these two spheres being, respectively, $4\pi r_1^2$ and $4\pi r_2^2$, the surface density of the intersections of the lines of force (recall that there are N of these being considered) on

these spherical surfaces will be $\frac{N}{4\pi r_1^2}$ and $\frac{N}{4\pi r_2^2}$ respectively. Evidently, the surface density of the lines of force will be larger on the first surface compared to that on the second surface, and the ratio of the two will be $\frac{1}{r_1^2} : \frac{1}{r_2^2}$.

In this context, note that the magnitudes of electric field intensities at distances r_1 and r_2 from a point source are also related in the *same* way, namely bearing a ratio $\frac{1}{r_1^2} : \frac{1}{r_2^2}$ to each other (check this out; refer to eq. 11-5c). In other words the surface density of lines of force is proportional to the electric field intensity. A similar conclusion applies more generally to an electric field set up by a number of source charges.

Fig. 11-18 depicts a part of a tube-like surface in an electric field that is entirely made up of lines of force of the field. This means that the line of force passing through any point on the surface of the tube tube lies entirely on its surface (indeed, this *defines* the surface of the tube). Such a structure is referred to as a *tube of force* in the electric field.

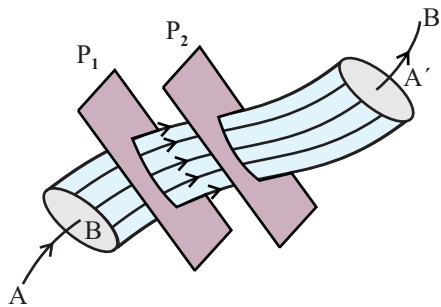


Figure 11-18: A part of a tube of force in an electric field; the line of force ABA'B' passes through the interior of this tube; P₁ and P₂ are two cross-sections of the tube.

Evidently, if a part of a line of force lies inside the tube, then that line of force cannot come out of it, because in order to do so, it has to intersect some other line of force lying on the surface of the tube (recall that two distinct lines of force cannot intersect each other). The line of force marked ABA'B' in fig. 11-18 is one such line of force that lies entirely inside the tube of force shown. If now one considers a *given number* (say, N) of such lines of force, then these will lie entirely within the tube under consideration. In the figure, the planes P₁ and P₂ are two cross-sections of the tube of force under

consideration, to which the lines of force are perpendicular, which means that these are two equipotential surfaces. Suppose that the area of cross section in the plane P_1 is larger than that in P_2 . This means that the given number (N) of lines of force pass through a larger area in P_1 as compared to that in P_2 . In other words, the *density* of these lines of force is smaller in P_1 compared to that in P_2 . Assuming that all the lines of force cross the planes P_1 and P_2 in the same direction (this requires that the part of the tube lying in between P_1 and P_2 contains no source points or neutral points), one can then make the statement that the electric field intensity on the section P_1 will be less than that on P_2 .

In mathematical terms, the integral $\int \mathbf{E} \cdot \hat{n} ds$, i.e., the *flux* of electrical intensity (see sec. 11.8.1), evaluated over any given cross-section of a tube of force, is independent of the cross-section chosen, where \hat{n} denotes the unit normal, along the direction of the lines of force, on such a cross-section. Considering a narrow tube, one has $E_1 \delta S_1 = E_2 \delta S_2$, where the subscripts 1 and 2 refer to any two cross-sections, δS_1 and δS_2 being the two areas of cross section. This means that the intensity on any given cross-section is inversely proportional to the area. For a given number of lines of force, this translates to the intensity being proportional to the density of the lines of force.

11.7.5 Separations between equipotential surfaces

Corresponding to the fact that the electric field intensity is monotonically related to the density with which the lines of force pass through a given cross section, one can correlate the intensity with the way a number of successive *equipotential surfaces* are arranged in the field. In the fig. 11-19 below, a number of equipotential surfaces are shown schematically, where AA' and BB' are two lines, not necessarily straight, that cut normally through these surfaces, i.e., in other words, these two are a pair of lines of force. Notice that the intersection points are closer to one another on AA' as compared to those on BB' , i.e., the successive equipotential surfaces lie closer to one another along AA' than along BB' .

This means that the magnitude of the rate of change of potential is larger along AA' than

along BB' , which implies that the component of intensity along AA' is larger than that along BB' (assuming that $V_1 > V_2 > V_3$, the directions of the intensity on the two lines of force are as shown by the arrow-heads). In other words, the closer the successive equipotential surfaces, the larger is the intensity.

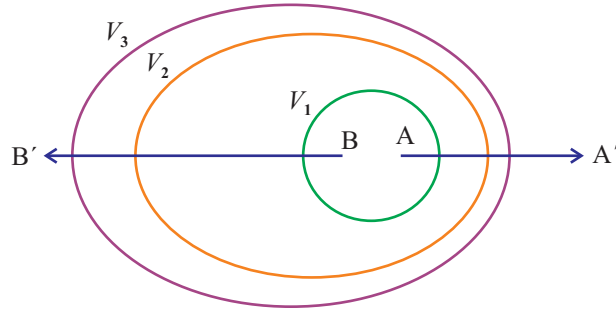


Figure 11-19: A set of equipotential surfaces in an electric field; AA' and BB' are two lines of force; the equipotential surfaces are closer to one another along AA' than along BB' , implying that the intensity is larger along AA' as compared to that along BB' .

11.8 Gauss' principle in electrostatics

The entire discussion in the present section resembles closely the corresponding discussion on Gauss' principle in gravitation (section 5.3).

11.8.1 Flux of electric field intensity

In fig. 11-20 below, S is a closed surface in an electric field, on which P is any chosen point. A small area around P lying on S has been shown in the figure, which can be assumed to be a plane one, lying in the tangent plane at P to the surface. Also shown are the outward drawn normal (PN) at P (i.e., the normal to the surface, pointing away from its interior) and the electric field intensity (E) at P , represented by the directed line segment PR .

If the area of the small element around P be δs and the unit vector along the normal PN be \hat{n} , then the vector area (see section 2.6.2) of the element will be $\vec{\delta s} = \delta s \hat{n}$. The

expression $\mathbf{E} \cdot \vec{\delta s}$ is then referred to as the *flux* of the electric field intensity through the element of area under consideration. Evidently this will be a small quantity in the present context. Denoting this by $\delta\Phi$ one has

$$\delta\Phi = \mathbf{E} \cdot \vec{\delta s} = E\delta s \cos \theta, \quad (11-47)$$

where θ is the angle between \mathbf{E} and \hat{n} .

Imagining the entire closed surface S to be divided into a large number of such small area elements, and obtaining the flux through each such element in the above manner one may finally work out the sum of all these small quantities so as to arrive at the *total* flux through the closed surface S :

$$\Phi = \sum \delta\Phi = \sum \mathbf{E} \cdot \vec{\delta s}, \quad (11-48)$$

where the summation is over all the small area elements on S . If, now, the area of each of the elements be assumed to be vanishingly small (i.e., to go to zero), then the flux reduces to what is referred to as the *surface integral* of the electric field intensity on S :

$$\Phi = \oint \mathbf{E} \cdot \vec{ds} = \oint E \cos \theta ds. \quad (11-49)$$

Here the symbol \oint indicates the surface integral on a closed surface.

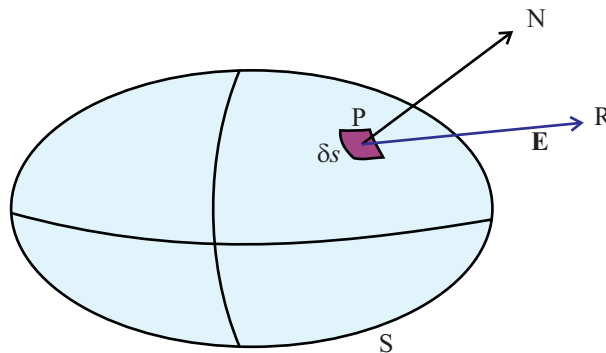


Figure 11-20: Illustrating the concept of electrical flux over a surface, the surface chosen being a closed one in the present instance; a small area element on the surface is shown at the point P, where the outward drawn normal PN and the electric field intensity \mathbf{E} along PR are also shown.

11.8.2 Gauss' principle

Evidently, the value of flux over any closed surface in an electric field will depend on the source charges responsible for the setting up of the field, where the closed surface may be an imagined one rather than the boundary surface of a material body. However, the relation between the source charges and the flux over a closed surface is a curious one, and forms the content of what is referred to as Gauss' principle in electrostatics.

Recall from eq. (11-6a) that the electric field intensity at any chosen point in an electric field is determined by the values and locations of *all* the source charges setting up the field. The *flux* over a closed surface, however, is determined *only* by the *total charge within* that closed surface. Denoting the latter by the symbol Q , the flux is given by

$$\oint \mathbf{E} \cdot d\vec{s} = \frac{Q}{\epsilon_0}, \quad (11-50)$$

where ϵ_0 stands for the permittivity of vacuum. This relation between the electrical flux over a closed surface and the total source charge located within that surface is the mathematical expression of Gauss' principle.

We are considering here, for the sake of simplicity, electrostatic fields set up in vacuum. A considerable number of basic concepts in electrostatics are conveniently formulated by referring to idealized situations where the source charges are imagined to be placed in vacuum. Features of electrostatic fields in conductors and dielectrics will be considered in sec. 11.10.

Notice that the electrical flux over a closed surface does not depend on the values and locations of the source charges external to that surface, nor does it depend on the locations of the source charges in the interior, depending instead on the total value of the interior charges. This is the content of Gauss' principle.

Gauss' principle is often a useful and convenient means for determining the electric field intensities in electric fields created by *symmetric distributions* of source charges.

One can also determine the intensity in an electric field by applying Coulomb's law in conjunction with the principle of superposition. However, Gauss' principle is at times a more convenient one in this respect. It has to be mentioned, though, that one needs both the Coulomb law and the superposition principle in establishing Gauss' principle, which is thus just a modified form of these two.

The mathematical expressions for Coulomb's law and Gauss' principle get modified in certain ways if the charges and field points happen to be located in a material medium.

11.8.2.1 Flux due to a single point charge

Consider, to start with, a single point charge q located at the point Q and a surface S , not necessarily a closed one (fig. 11-21) in the electric field produced by q . Imagining the surface S to be divided into a large number small surface elements, and any one of these elements, say, δs , the flux of electric field intensity through this area element is given by $E\delta s \cos \theta$ in the notation of fig. 11-20 and eq. (11-47). Denoting by r the vector extending from Q up to the point P in the area element considered, one has, $E = \frac{1}{4\pi\epsilon_0} \frac{q}{r^2}$ so that the expression for the flux can be written as

$$\delta\Phi = \frac{q}{4\pi\epsilon_0} \frac{\delta s \cos \theta}{r^2}. \quad (11-51)$$

11.8.2.2 Solid angle

In this expression, θ stands for the angle between the line joining Q and P , and the normal to the surface S at the point P . The small quantity given by the expression $\frac{\delta s \cos \theta}{r^2}$ in the above equation is referred to as the *solid angle* ($\delta\Omega$) subtended at Q by the area element δs :

$$\delta\Omega = \frac{\delta s \cos \theta}{r^2}. \quad (11-52)$$

Imagining the points on the boundary of the area element δs to be connected to the point Q , $\delta\Omega$ gives a measure of the opening of the cone formed by these connecting lines.

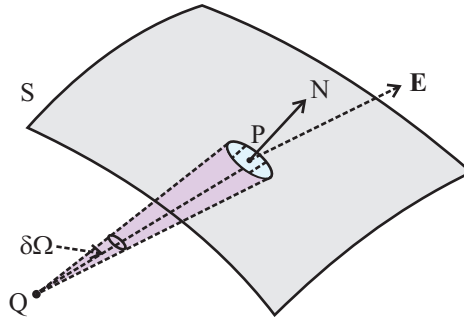


Figure 11-21: The flux through the surface S of a single point charge at Q ; a small element of area δs around the point P is shown; the electric intensity at P due to the point charge is along QP , which makes an angle θ with the normal PN to S at P ; imagining the points on the boundary of the area element δs to be connected by straight lines to Q , the quantity $\frac{\delta s \cos \theta}{r^2}$ occurring in eq. (11-51) is the solid angle subtended by δs at Q , which represents the opening of the cone formed by these connecting lines.

If, now, the elementary solid angles subtended at Q by all the area elements making up the surface S are summed up, one gets the solid angle (Ω) subtended by S at Q . In the limit of δs going to zero, the sum reduces to an integral defining the solid angle

$$\Omega = \int_S \frac{ds \cos \theta}{r^2}. \quad (11-53)$$

Imagining once again the points on the boundary of S to be connected by line segments with Q , the solid angle gives a quantitative measure of the opening of the cone formed by these connecting lines.

The solid angle subtended by a surface S at a point Q , as defined above, is characterized by the following remarkable property. Consider a second surface S' , as in fig. 11-22 such that the cone formed by lines connecting boundary points of S' with Q is the same as the cone obtained with S (see fig. 11-22). *The solid angles subtended at Q by S and S' will then be equal.*

Making use of this property of the solid angle, one can now arrive at the following two equally remarkable conclusions: (a) *the solid angle subtended by a closed surface at any interior point is 4π regardless of the location of the point in the interior of the surface,* and (b) *the solid angle subtended by a closed surface at an exterior point is zero.*

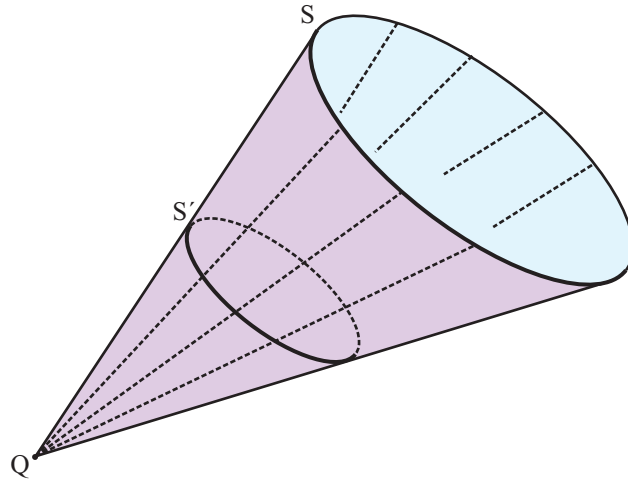


Figure 11-22: Illustrating the equality of two solid angles; the solid angles subtended at Q by the surfaces S and S' are equal.

The first of these two results is obtained by noting that the solid angle subtended at the point Q by the closed surface S (fig. 11-23(A)) is the same as the solid angle subtended by a sphere drawn around Q as centre (reason this out). Considering any point P on the sphere, one has $\theta = 0$, $r = R$, where R stands for the radius of the sphere. The solid angle is then given by

$$\Omega = \frac{1}{R^2} \int ds = 4\pi, \quad (11-54)$$

since the surface area of the sphere is $4\pi R^2$.

For an exterior point, on the other hand, one can imagine an almost closed surface S with a small opening in it as in fig. 11-23(B), in which case the cone referred to above is obtained by connecting the boundary points on this small opening with the point Q . In the limit of the opening being made to shrink to a point, S reduces to a closed surface and the solid angle reduces to

$$\Omega = 0. \quad (11-55)$$

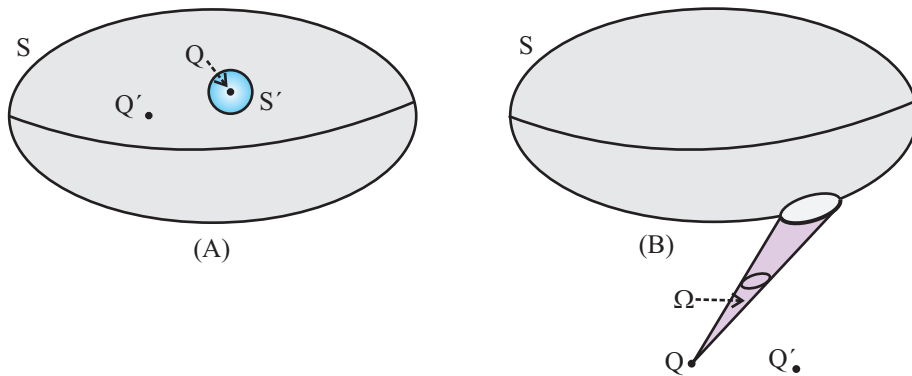


Figure 11-23: Illustrating the solid angle subtended by a closed surface at (A) an interior point, and (B) an exterior point; in (A), the solid angle subtended at Q by S is the same as that subtended by the spherical surface S' ; in (B) the closed surface is replaced with an almost closed one with a small opening in it; the solid angle Ω reduces to zero as the opening is made to shrink to a point.

What is important to note is that the solid angle subtended by a closed surface at an interior or an exterior point does not depend on the actual location of the latter. In other words, considering any other point, say, Q' in fig. 11-23(A), interior to the closed surface S , the solid angle will still be 4π , and similarly, the solid angle subtended at the exterior point Q' in fig. 11-23(B) will still be zero.

The solid angle subtended by a surface S at a point lying *on* the surface is 2π (reason this out).

11.8.2.3 Gauss' principle: derivation

Looking back at equations (11-51) and (11-52), the flux through an elementary area δs on a closed surface S due to a single point charge q located at Q is seen to be

$$\delta\Phi = \frac{q}{4\pi\epsilon_0}\delta\Omega, \quad (11-56)$$

where $\delta\Omega$ stands for the solid angle subtended at Q by the elementary area under consideration. Thus, summing up over all the small area elements making up the closed surface S , the flux through the entire closed surface will be given by $\frac{q}{4\pi\epsilon_0}\Omega$, where the solid angle Ω is 4π or 0 according as the charge q is located in the interior of S or is exterior to it.

I can now outline for you the derivation of Gauss' principle in electrostatics which, with necessary modifications, applies to gravitation as well (refer back to sec. 5.3.2). Consider, then, a closed surface S in the field of a number of charges, say, q_1, q_2, \dots, q_N , of which some may be located in the interior of S , the others being exterior to S .

The total flux through S due to all these source charges is, by the principle of superposition, the sum of the fluxes due to the charges considered severally (reason this out). According to the above result obtained for a single source charge in the interior of or exterior to a closed surface, the source charges exterior to S do not contribute to the flux while a source charge, say, q_i (where i belongs to the set $\{1, 2, \dots, N\}$) interior to S contributes a flux $\frac{q_i}{\epsilon_0}$.

The total flux through S is then given by $\frac{1}{\epsilon_0}$ times the sum of all the source charges interior to it, which completes the required derivation, giving (11-50), where the left hand side is, by definition, the electrical flux through the closed surface under consideration.

11.9 Applications of Gauss' principle

As already mentioned, Gauss' principle in electrostatics, expressed by eq. (11-50) resembles closely the corresponding principle in gravitation, expressed by eq. (5-16), where a replacement $\mathbf{I} \rightarrow \mathbf{E}$, $-4\pi GM \rightarrow \frac{Q}{\epsilon_0}$ reduces the latter to the former. Hence, results derived from this principle in gravitation and in electrostatics bear a close resemblance to one another.

11.9.1 A charged spherical conductor

For instance, considering a spherical charge distribution (with the origin chosen at the center of the distribution) with a total charge Q , the result (5-17b) gives the electric field intensity at a point \mathbf{r} as

$$\mathbf{E}(\mathbf{r}) = \frac{q(r)}{4\pi\epsilon_0 r^3} \mathbf{r}, \quad (11-57)$$

where $q(r)$ stands for the charge lying within a spherical Gaussian surface of radius r .

In particular, let a charge Q be given to a spherical conductor of radius R imagined to be isolated from all other bodies, where the center of the sphere is chosen as the origin. As will be seen in 11.10.3, the charge density everywhere in the interior of a conductor is zero, and any charge given to it resides on its surface. Due to the spherical symmetry of the conductor, which is assumed to be isolated from all other bodies, the charge distribution on the surface of the conductor will be a spherically symmetric one.

We can now make use of the result (11-57) and conclude that the intensity at a point located outside the conductor at a distance r ($r > R$) is given by

$$\mathbf{E}(\mathbf{r}) = \frac{Q}{4\pi\epsilon_0 r^3} \mathbf{r}, \quad (11-58)$$

which is identical to the intensity due to a point charge Q placed at the origin. If, on the other hand, the point under consideration lies within the sphere, then the intensity at that point is seen to be zero.

Knowing the intensity due to the charge on the spherical conductor at any point P at a distance r from the center O, one can evaluate the potential $V(r)$ at P by evaluating the integral $-\int \mathbf{E} \cdot d\mathbf{r}$ (refer to eq. (11-7)) along any path connecting a chosen reference point and the point P. In the present instance, the source charge on the conductor being contained in a finite region of space, the reference point may be chosen to be at an infinitely large distance (see sec. 11.4.4). The path may be chosen for the sake of convenience to lie along the line OP extended towards infinity, which gives

$$V(r) = \frac{q}{4\pi\epsilon_0 r}, \quad (11-59a)$$

if the point P lies outside the spherical surface of the conductor, and

$$V(r) = \frac{q}{4\pi\epsilon_0 R}, \quad (11-59b)$$

if P lies in the interior of the surface. The result (11-59a) is derived as in the case of

eq. (5-18b) while (11-59b) is obtained by noting that the intensity in the interior of the sphere being everywhere zero, the potential has to be constant in the interior, its value being the potential at the surface of the conductor.

In this context, recall the result relating to the gravitational force between two rigid non-overlapping spherical bodies, considered in sec. 5.3.3.3, which was seen to be equal to the force between two equivalent point masses imagined to be located at the centres of the respective spheres. A corresponding result does *not*, however, hold in the electrostatic situation. Considering two spherical conductors, for instance, each with a charge given to it, the mutual interaction between the charges will result in a charge distribution which will no longer be a spherically symmetric one. The derivation of the force between the two charged spherical conductors will thus involve more detailed considerations, which I do not enter into (indeed, this problem admits of no simple solution for arbitrary radii of the two spheres and arbitrary separation between their centers).

11.9.2 A spherically symmetric charge distribution

As mentioned above, these results relating to a charged spherical conductor are special instances of those pertaining to a spherically symmetric charge distribution (see eq. (11-57)). The field intensity and potential due to such a spherically symmetric charge distribution can be obtained in a manner entirely analogous to the corresponding results for the gravitational intensity and potential due to spherically symmetric body (sec. 5.3.3). Considering, for instance, a spherical body of radius R containing a charge Q distributed with spherical symmetry, the field intensity $E_1(r)$ at an external point at a distance r ($r > R$) from the centre, which is directed radially, is given by the expression

$$E_1(r) = \frac{Q}{4\pi\epsilon_0 r^2}, \quad (11-60)$$

where it is assumed that the medium outside the body is free space. The intensity (once again directed radially) at an internal point at a distance r ($r < R$), on the other hand,

is given by the expression

$$E'(r) = \frac{q(r)}{4\pi\epsilon_0 r^2}, \quad (11-61)$$

where $q(r)$ stands for the charge within an imagined sphere of radius r concentric with the given body (thus, $q(R) = Q$). This expression (which is a re-statement of eq. (11-57)), however, differs from the actual field in the body and is referred to as the ‘vacuum field’. The actual field depends on the material of the body under consideration. For a large class of bodies, the relevant material property is expressed in terms of a single scalar quantity referred to as the *relative permittivity* (ϵ_r , see sec. 11.10.5.2). The field strength is then given by the expression

$$E_2(r) = \frac{q(r)}{4\pi\epsilon_r\epsilon_0 r^2}. \quad (11-62)$$

Having obtained the field intensity $E(r)$ at any point, external or internal with reference to the body, and noting that the intensity is directed radially, the potential $V(r)$ at any point at a distance r from the center can be obtained from the expression

$$V(r) = - \int_{\infty}^r E(r') dr', \quad (11-63)$$

where the integral is to be performed along a radially directed straight line (reason this out).

For a *uniformly charged* spherical body of radius R , this works out to

$$V(r) = \frac{Q}{4\pi\epsilon_0 r} \quad (r > R), \quad (11-64a)$$

for a point at a distance r from the center, exterior to the sphere, and

$$V(r) = \frac{Q}{4\pi\epsilon_0\epsilon_r R^3} \frac{3R^2 - r^2}{2} \quad (r < R), \quad (11-64b)$$

for an interior point at a distance r (derive these results, following the approach leading to the results (5-24a) and (5-24b)).

Problem 11-9

A spherical raindrop carrying a charge $q = 20 \times 10^{-12}$ C, distributed with spherical symmetry in the drop, has a potential $V = 400$ V on its surface. Find the radius R of the drop. If two such drops, identical in all respects, combine to form a single drop, with the charge once again distributed with spherical symmetry, what will be the potential at its surface?

Answer to Problem 11-9

HINT: Assuming the potential at infinity to be zero, the potential on the surface of the drop is $V = \frac{q}{4\pi\epsilon_0 R}$, i.e., the required radius is $R = \frac{20 \times 10^{-12}}{4 \times 3.14 \times 8.85 \times 10^{-12} \times 400}$ m = 4.5×10^{-4} m (approx). When two such drops combine, the charge will be $q' = 2q$, and the radius will be $R' = 2^{\frac{1}{3}}R$ (reason this out). Thus, the potential will now be $V' = \frac{2}{2^{\frac{1}{3}}}V = 635$ V (approx).

11.9.3 Potential energy of a uniformly charged sphere

Considering any given system of charges, one can associate a certain amount of energy with the system, referred to as its electrostatic self energy, which is defined as the energy required to set up the charges at their given positions, starting from a configuration where the charges are separated from one another by infinitely large distances, this energy, in turn, being the work done against the electrostatic forces between the particles brought into play during the process of assembling the charges at their given positions. It is assumed that, during the process of assembly, the charges are moved quasi-statically, i.e., with vanishingly small kinetic energy.

Imagine, for instance, such a process of assembling a uniformly charged sphere of radius R , with a charge Q uniformly distributed in it. The charge density (refer to formula (11-10)) is thus $\rho = \frac{Q}{\frac{4}{3}\pi R^3}$. Suppose that, at any intermediate stage of the process of assembly, the sphere has a radius r , when the radius is increased by δr , with a charge $\delta q = 4\pi r^2 \rho \delta r$ brought on to it from an infinite distance. The potential at the surface of the sphere at this stage being $V_1(r) = \frac{\frac{4}{3}\pi r^3 \rho}{4\pi\epsilon_0 r}$ (refer to eq. (11-64a), where the potential is a continuous function of the position), the energy required for bringing in the additional

charge is $V_1(r)\delta q = 4\pi r^2 \rho V_1(r)\delta r$. The total energy is then obtained by a process of summation which reduces to an integration as δr is made infinitesimally small. On working out the integral, one gets the required energy as

$$W = \frac{1}{4\pi\epsilon_0} \frac{3}{5} \frac{Q^2}{R}, \quad (11-65)$$

(check this out).

In general, the electrostatic energy of a system of point charges is given by the expression (11-14) where V_i ($i = 1, \dots, N$) denotes the potential at the position of the charge q_i belonging to the system, due to the other charges in it. Here there is no explicit reference to the process of assembling the charges, i.e., in this expression, V_i is the potential at the location of the charge q_i when the charges *have already been assembled*.

In the case of the uniformly charged sphere, the charge in a thin shell of inner and outer radii r and $r + \delta r$ is δq given above, while the potential $V(r)$ at the distance r from the centre is given by the formula (11-64b). One thereby obtains the electrostatic energy of the sphere as $W = \frac{1}{2} \sum \delta q V(r)$. This reduces to an integral for vanishingly small δr , giving

$$W = \frac{1}{2} \int_0^R 4\pi r^2 \rho \frac{Q}{4\pi\epsilon_0 R^3} \frac{3R^2 - r^2}{2} dr.$$

On evaluating this integral, one arrives at the same result for W as in (11-65), where it is assumed that the charge elements making up the spherical distribution are all located in vacuum (check this out).

11.9.4 An infinitely long cylindrical conductor

Consider a uniform charge distribution on the surface of a long cylinder where the length of the cylinder is so large (compared to its radius, say, R) that it can be taken to be effectively infinite. Let P be any point located outside the cylinder at a distance r (fig. 11-24). Imagine a Gaussian surface, coaxial with the surface of the conductor, and passing through the point P. Owing to the symmetry of the problem, the electric field intensity at P will be directed along the line NP, where N is the foot of the perpendicular

dropped from P on the axis of the cylinder and, moreover, the magnitude of the intensity will be the same as that at any other point, say Q, on the Gaussian surface.

If E denotes the magnitude of the intensity at P, and λ the charge per unit length of the cylinder then, considering a cylindrical Gaussian surface of unit length one gets, from the result expressed by eq. (11-50), applied to the situation under consideration,

$$2\pi r E = \frac{\lambda}{\epsilon_0}, \quad (11-66)$$

(check this formula out).

In other words, the magnitude of the intensity at the point P is given by

$$E = \frac{\lambda}{2\pi\epsilon_0 r}, \quad (11-67)$$

the intensity being directed along NP.

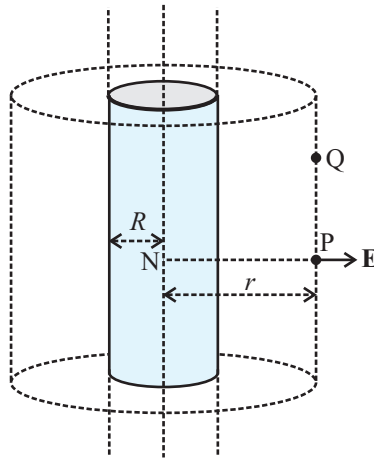


Figure 11-24: Field due to a uniform charge distribution on the surface of a cylinder, assumed to be infinitely long; the field intensity at a point P is along NP, where N is the foot of the perpendicular dropped from P on the axis of the cylinder, and its magnitude is the same for all points on the Gaussian surface which in this case is a coaxial cylinder through P.

If, on the other hand, the point P is located in the interior of the cylindrical surface, then the intensity at P is zero (see sec. 11.10.3).

The potential at the point P is once again obtained by evaluating the integral $-\int \mathbf{E} \cdot d\mathbf{r}$ along a convenient path, where we can take the path to lie along the line NP. However, the charge on the cylinder in the present problem is *not* confined within a finite region of space, and hence the ‘point at infinity’ is not an appropriate reference point here. Instead, let us assume that the reference point is located at a distance r_0 from the axis of the cylinder (shifting the reference point parallel to the axis will not change the result we derive), in which case one obtains

$$V(r) = -\frac{\lambda}{2\pi\epsilon_0} \ln\left(\frac{r}{r_0}\right) \quad (r > R), \quad V(r) = -\frac{\lambda}{2\pi\epsilon_0} \ln\left(\frac{R}{r_0}\right) \quad (r < R), \quad (11-68)$$

(check these results out).

11.9.5 An infinitely extended planar sheet of charge

Consider now an infinitely extended planar sheet of charge (fig. 11-25), the surface density (say, σ) of charge in the sheet being uniform (the surface charge density, i.e., the charge per unit area of a surface, is defined in formula (11-71) below). The symmetry of the problem implies that the field intensity at any point P is the same everywhere, regardless of the distance of P from the planar sheet, being directed along a line perpendicular to the plane (S) of the sheet (reason this out). Let this intensity, along the unit vector \hat{n} directed away from the plane S (on either side of it) be E .

Let us consider a Gaussian surface of the shape of a cylinder, with one end face of the cylinder passing through P, and with the axis of the cylinder perpendicular to the plane S, the other end face of the cylinder being on the other side of S, as shown in fig. 11-25. The charge enclosed by the Gaussian surface is $\sigma\delta S$ where δS stands for its area of cross section.

As regards the flux through the Gaussian surface, the contribution by the part of the surface perpendicular to the planar sheet S (this part of the Gaussian surface is denoted by Σ in fig. 11-25) is zero since the intensity at any point on Σ is tangential to Σ itself, implying that $\cos\theta = 0$ in eq. (11-52) for any and every small part of Σ . What remains

of the flux through the Gaussian surface is the contribution through the two end faces. Since the unit normal to either end face is directed away from the planar sheet S, its contribution to the flux is seen to be $E\delta S$.

Thus, eq. 11-50 assumes the form

$$2E\delta S = \frac{\sigma\delta S}{\epsilon_0}, \quad (11-69)$$

giving the required expression for the electric field intensity due to the charged planar sheet as

$$E = \frac{\sigma}{2\epsilon_0}. \quad (11-70)$$

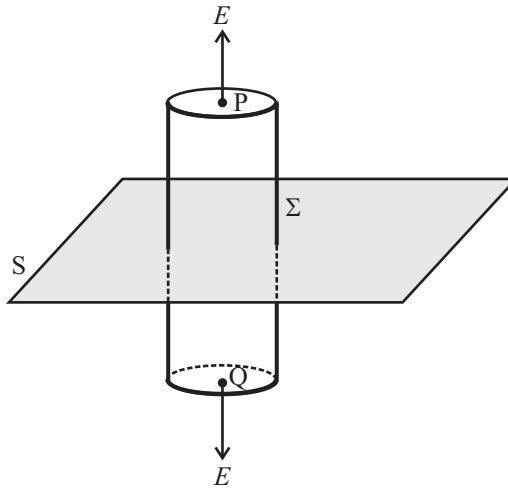


Figure 11-25: Field due to infinitely extended planar sheet of charge (S), with uniform surface charge density; the Gaussian surface is a cylindrical one with area of cross-section δS ; \hat{n} denotes the unit vector normal to the surface S on either side of it; Σ denotes the part of the Gaussian surface perpendicular to S; for any point on Σ , the direction of the electric field intensity is perpendicular to the normal at that point.

If a charge is given to a conducting planar sheet, it does not distribute itself uniformly on the sheet. In the case of a thin circular disk, for instance, it is found that the charge density is minimum at the center, increasing to large values near the rim of the disk. However, if two such conducting sheets of large area are brought close to each

other, with their planes parallel, forming a *parallel plate capacitor* (see sec. 11.11.8) the charge given to either sheet distributes itself with approximately uniform density on it (the charge densities on the two faces of the disk being, however, different; see sec. 11.11.8.1).

For a uniformly charged sheet of large but finite area, the field remains uniform to a good degree of approximation (being given by the expression (11-70)) only up to a certain distance from the conductor, beyond which the field intensity falls off with the distance.

11.10 Conductors and dielectrics

11.10.1 Free and bound electrons

Till now, it has been assumed that all source charges are located in vacuum and the electric fields are also set up in vacuum rather than in material media. We now turn to a consideration of electric fields in material media. From the point of view of electrical behavior, such media can be classified into two broad groups, namely, *conductors*, and *dielectrics*. The principal distinguishing feature for these two types of media is that, while an electric *current* can be set up in a conducting medium by an electric field, no such current can be produced in a dielectric unless the electric field intensity in it is very high.

This difference between conductors and dielectrics is related to their internal structural characteristics. A conducting material contains a large number of ‘free’ electrons, i.e., ones that are not bound to specific atoms or molecules within it. Consequently, even a weak electric field can set up a current in the conductor that can be sustained with the help of appropriate arrangements. In contrast, almost all electrons in a dielectric material are bound to specific atoms or groups of atoms, and are not capable of producing an electric current when an electric field is set up in it. Such materials are therefore electrical *insulators*.

1. A sufficiently strong electric field can be made to tear away the bound electrons in a dielectric material from the atoms to which these are attached. However, the effects of such strong electric fields will not be considered here.

While conducting materials are, as a rule, crystalline in structure, an insulator can be either crystalline or amorphous. The electrical properties of a material depend on a number of basic features of the way their constituent atoms and molecules are held together. However, I will not enter into these considerations here (for a brief discussion, see sections 19.2.1, 19.2.6).

2. In this context, I need to refer to what are known as *semiconductors*. These are similar in nature to conductors in that these can carry currents generated by electric fields. But the conductivity of a semiconductor being small compared to that of a conductor, the former can be looked at as a dielectric from the point of view of electrostatics.

11.10.2 Electric field intensity and charge density within a conductor

Looked at from the point of view of electrostatics, the fundamental difference between conductors and dielectrics pertains to the fact that *the electric field intensity at every point within a conductor in the condition of static equilibrium is zero*.

Here the term ‘static equilibrium’ means a condition where the charge density, electrical potential, and electric field intensity at every point in an electric field are time independent, and there is no current flowing in the material medium under consideration. This is the type of situation one considers in electrostatics.

In electrostatics, one other important characteristic property of a conductor is that, along with the electric field intensity, the *charge density is also zero everywhere within a conductor*.

Imagining a given region of space to be divided into a large number of small volume elements, the ratio of the charge contained within any one of these elements and the volume of that element is referred to as the ‘charge density’ in it. Put differently, if the

charge within a small volume δV around a point P is δq , then the charge density (say, ρ) at P is given by the ratio $\frac{\delta q}{\delta V}$ in the limit $\delta V \rightarrow 0$ (refer to sec. 11-10).

If the point P is situated in the interior of a material medium, then any volume element around that point will in general contain both positive charges in the nuclei of the material and negative charges in the form of electrons. A charge density $\rho = 0$ at P means that the positive and negative charges within the volume element V cancel each other.

What can be the reason underlying the fact that the charge density everywhere in the interior of a conductor is zero under electrostatic conditions? The explanation lies in the *mobility* of the free electrons in a conducting material I have alluded to above. Due to this mobility, the free electrons start moving around in even a weak electric field set up in the conductor. The electrons keep on moving and the charge density keeps on changing as long as the electric field intensity remains non-zero in the conducting medium.

In the end, a static condition prevails when the intensity as well as the charge density reaches zero value everywhere (a non-zero charge density would imply that there remains an electric field in the material, due to which the free electrons would continue to move around and the static condition would not hold). This, in reality, is a self-adjusting process that terminates when the electrostatic condition is reached. In a dynamic condition, however, when the electrical potential and intensity in the medium change with time, neither the charge density nor the electric field intensity need be zero within a conducting medium.

11.10.3 Conductor: surface charge density.

The fact that the electric field intensity in the interior of a conductor is zero under electrostatic conditions, implies that *the potential has to be uniform throughout its volume*, because the electric field intensity in any given direction is nothing but the rate of change of potential with distance along that direction (with a negative sign).

Evidently, the constant value of the potential inside the conductor implies the same constant value on its *boundary surface* as well. In other words, the boundary surface of a conductor is an *equipotential* one.

As I have already mentioned, the charge density in the interior of a conductor is zero under electrostatic conditions. This is a consequence of the self-adjusting process mentioned above in which the local number densities of the free electrons (i.e., the number densities at various different points in the medium) keep on changing till the intensity goes to zero and a static condition prevails. In other words, *any charge given to a conductor resides on its boundary surface*, the charge density everywhere in the interior of the conductor being zero.

Imagine a small element of surface around any chosen point P on the surface of the conductor (see fig. 11-26; this can be looked upon as part of a plane surface). If the area of this small element be δA and the charge in this element be δq , then the ratio

$$\sigma = \frac{\delta q}{\delta A}, \quad (11-71)$$

is termed the *surface charge density* at the point P (strictly speaking, one should consider here the limiting case when δA is vanishingly small). The unit of surface density of charge is $\text{C}\cdot\text{m}^{-2}$.

The term ‘density’ is used in physics in various different contexts. In the first place, one may refer to densities of various different physical quantities like, for instance, mass or charge. Although it is the mass-density that is commonly referred to as, simply, density, quantities like charge density or number density may also be relevant in certain contexts. Moreover, the density under consideration may be defined with reference to unit volume, unit surface area, or even unit length. For instance, one may refer to volume density of charge, surface density of charge, or linear density of charge. Looked at from this point of view, the charge density defined at any point

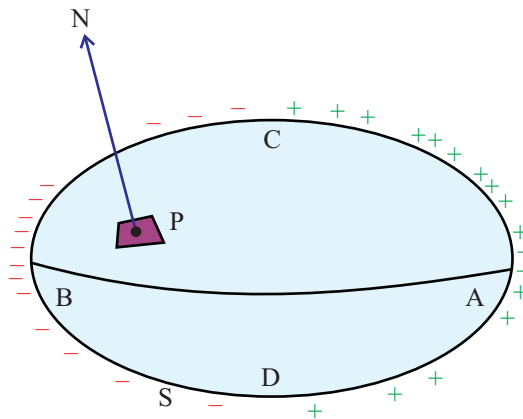


Figure 11-26: Illustrating the concept of surface charge distribution and surface density of charge on the boundary surface of a conductor; S is the boundary surface of a conductor, on which P is any chosen point; the charge in a small element of area around P determines the surface charge density at P; the distribution of charge on the surface is illustrated schematically; the '+' symbols denote positive surface density of charge while the '-' symbols denote negative surface density; the magnitude of surface density is, in general, large at sharply pointed regions like A and B, and small at flat regions like C and D; the electric field intensity at any point like P is directed along the normal PN and may point either outward or inward.

on the surface of a conductor is a surface density, while the number density of free electrons at any point in a conductor is a volume density, the unit for which is m^{-3} .

The surface charge densities at the various points on the boundary surface of a conductor may be, in general, different (as shown schematically in fig 11-26). Indeed, the free electrons distribute themselves in a self-adjusting process in such a way that, as the electrostatic condition is reached, both the volume charge density and the electric field intensity in the interior of the conductor become zero, and it is this process that determines the charge density distribution on the surface. In other words, the net or total charge of the conductor gets distributed on its surface in just the right manner so as to reduce the intensity at every interior point to zero.

The fact that the volume charge density everywhere in the conductor is zero, is a logical consequence of the fact that the electric field intensity is zero, which can be seen by making use of Gauss' principle. However, the *reverse* reasoning is not valid. Given the statement that the volume charge density is zero, one cannot conclude that the electric field intensity has to be zero. Such a conclusion would be valid if, in addition, it is

assumed that the potential everywhere on the surface of the conductor is constant, which it actually is.

Analogous to the fact that the volume density of charge in the interior of a conductor is zero while the surface density on the boundary surface need not be zero, the electric field intensity at the boundary surface may also have a non-zero value unlike the intensity at interior points. However, *the field intensity at any point on the boundary surface, has to be directed along the normal to the surface at that point*, either toward the exterior or toward the interior of the conductor (i.e., either from P to N or from N to P in fig. 11-26 at the point P). It is not difficult to see why this should be so.

As I have stated above, the lines of force in an electric field are everywhere normal to the equipotential surfaces. Since the boundary surface of a conductor is an equipotential surface, the lines of force, and hence the intensity vectors have to be everywhere normal to the boundary, in the direction from a higher to a lower potential. Indeed, the surface being an equipotential one, the rate of change of potential along any direction *on* the surface has to be zero, implying that the component of electric field intensity along any such direction is zero.

If the component of intensity along any direction on the surface were non-zero, a force would have been there acting on the free charges located on the surface which would then have started moving around, thereby violating the electrostatic condition. There would then have taken place a redistribution of charges till the boundary surface would have been reduced to an equipotential one.

11.10.3.1 Field intensity on the surface of a charged conductor

There exists a relation between the surface charge density (say, σ) at any point P (refer to fig. 11-26) on the boundary surface of a conductor and the electric intensity E_n at that point. Here E_n is defined as the intensity along the outward drawn normal at P (i.e., from P to N in the figure; an intensity directed from N to P would then correspond to a negative value of E_n ; indeed, the field intensity at any point on the surface of a conductor is a one dimensional vector, and can be completely described by a signed

scalar, where the sign is taken to be positive for a field intensity directed outward).

This relation between σ and E_n can be expressed as

$$E_n = \frac{\sigma}{\epsilon_0}, \quad (11-72)$$

where ϵ_0 stands for the permittivity of free space. One concludes from this equation that a positive surface charge density corresponds to an outward intensity and a negative surface charge density to an inward one.

In order to see why this should be so look at fig. 11-27 where P is a point on the surface S of a conductor, at which the surface charge density is, say, σ . In the figure, the region below S is occupied by the material of the conductor while that above S is assumed to be vacuum. Consider a cylindrical Gaussian surface with end faces Σ_1 and Σ_2 on either side of S, being parallel to it, and with the remaining surface Σ perpendicular to S.

Assuming that the height of the Gaussian surface (i.e., its dimension in a direction perpendicular to S) is vanishingly small, the field intensity at any point on Σ_1 can be assumed to be E_n along \hat{n} , the unit normal to S in a direction away from the body of the conductor. Thus, the contribution of this part of the Gaussian surface to the electrical flux is seen to be $E_n \delta S$, where δS is the area of Σ_1 .

The contribution of Σ_2 to the flux is zero since the field intensity inside the body of the conductor is zero. The contribution of the remaining part (Σ) to the flux is also zero since the field intensity at any point on Σ , being along \hat{n} , is perpendicular to the normal to Σ .

In other words, the total electrical flux through the Gaussian surface is seen to be $E_n \delta S$. The charge enclosed within this surface, on the other hand, is $\sigma \delta S$. The result (11-72) then follows by Gauss' principle.

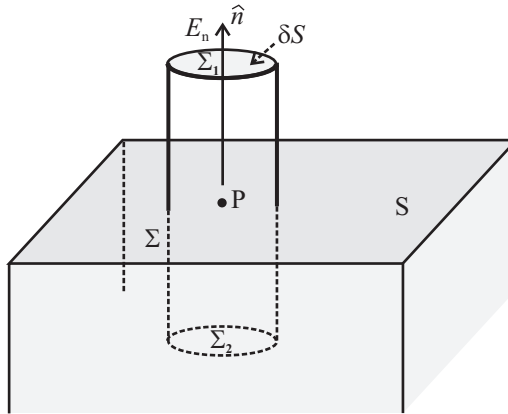


Figure 11-27: A cylindrical Gaussian surface at a point P on the surface S of a conductor, consisting of end faces Σ_1 , Σ_2 , along with the surface Σ whose height is assumed to be vanishingly small; the area of cross section of the cylinder δS is also small; the field intensity at P above the surface of the conductor is along \hat{n} while that inside the conductor is zero; the intensity E_n is related to the surface charge density σ at P as in eq. (11-72).

11.10.3.2 Force on the surface of a charged conductor

Because of the charge on the boundary surface of a conductor and the electric field intensity, given by (11-72), directed normally to the surface at any point (say, P) on it, there arises a mechanical force on the conductor acting on any small element of area around P. The force per unit area on the surface around P can be worked out by making use of the result in sec. 11.5. In this case, the total field strength E_- at any point within the conductor close to the point P on the surface is zero, while the field strength E_+ at an external point close to the surface is given by the right hand side of formula (11-72), the discontinuity in the total field being caused by the 'self-field' as explained in sec. (11.5). The average field strength across the surface of the conductor is thus $E_{av} = \frac{\sigma}{2\epsilon_0}$, where σ stands for the surface density of charge at the point P. Hence, the surface density of force at the point P under consideration is given by (refer to formula (11-23))

$$f = \frac{\sigma^2}{2\epsilon_0}, \quad (11-73)$$

where the force acts along the outward drawn normal to the surface at P, regardless of the sign of σ .

11.10.4 Accumulation of charge at sharp points

Suppose that a spherical conductor has been given Q amount of charge and is kept at a large distance from other bodies so that it is isolated from the effects of electric fields caused by other charges. What will then be the distribution of the charge on the surface of the conductor (recall that all the charge of a conductor has to reside on its surface under electrostatic conditions) ? Evidently, the spherical symmetry of the conductor and the fact that there are no external effects influencing the charge distribution, imply that the charge will distribute itself *uniformly* on the spherical surface. Indeed, one can make use of Coulomb's law and the superposition principle to show (I will skip the derivation for the sake of brevity) that a uniform charge distribution in this case implies that the surface of the conductor will be an equipotential one.

Now look at fig. 11-26 where an ellipsoidal conductor is shown, and assume that a charge Q is given to it, while ensuring once again that it is removed away from the possible influence of other charges and fields. How will the charge be distributed on the surface of the conductor in *this* case? The answer to this is no longer easy since there is no obvious symmetry here to help us find it. However, mathematical calculations can once again be invoked to work out the required charge distribution. Without going into these calculations, I will state an important conclusion resulting from these: the surface charge density will be greater in the regions A and B of the conductor compared with the regions C and D.

The section of the ellipsoid shown in fig. 11-26 is an ellipse. The curvature of this ellipse at A and B is larger than that at C and D. The curvature at A and B increases, i.e., these two regions become more and more sharp while the regions near C and D become flatter, as the minor axis along CD is made progressively smaller compared to the major axis AB of the ellipse. The mathematical calculations referred to above lead to the result that the sharper the regions A and B are, the greater is the surface charge density at these regions compared to the flat regions at C and D.

An *ellipsoid* is a three dimensional figure characterized by the property that every

plane section of it is an ellipse. The *curvature* at any point of a plane curve such as an ellipse is an indicator of how sharp or flat the curve is near that point. I need not enter here into a mathematical definition of this concept.

The accumulation of charge near sharply pointing regions on a conductor can be understood from fig. 11-28 where a set of equipotential surfaces in the region of space surrounding the ellipsoid is shown. Observe that the equipotential surfaces close to the boundary surface of the conductor are of a shape similar to the ellipsoid while those far away from it are nearly spherical.

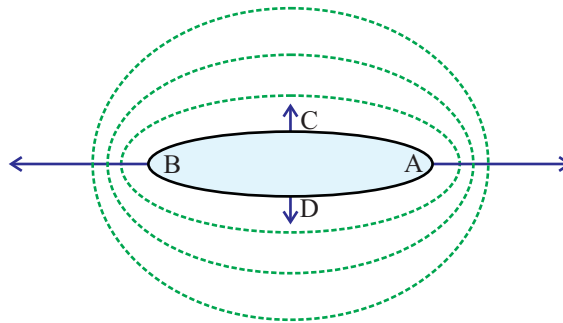


Figure 11-28: Explaining the accumulation of charge near a region of large curvature on a conductor; an ellipsoidal conductor is shown with a large curvature at A and B and a relatively smaller curvature at C and D; a set of equipotential surfaces around the conductor is also shown, where it is seen that the successive equipotential surfaces are closer to one another near A and B and separated with larger spacings in between, at C and D; this implies a larger electric field intensity at A and B as compared to that at C and D.

Far away from the conductor, the electric field resembles that of a point charge (a monopole) placed at the center of the ellipsoid.

One then observes that the successive equipotential surfaces are close to one another near the sharply projecting regions A and B while these are separated by relatively large spacings near the flat regions C and D. Since the electric field intensity is the rate of change of potential, one concludes that the intensity is larger at A and B as compared to the intensity at C and D. Finally, the relation of proportionality (eq. (11-72)) between the intensity and surface charge density on the surface of a conductor implies that the

charge density will be larger at sharply pointing regions as compared to the density at flat regions on the conductor.

This conclusion holds generally for a conductor which is not necessarily of an ellipsoidal shape. The fact that charges tend to accumulate with comparatively larger surface density near a sharply pointing region on a conductor leads to a number of consequences and applications. In fig. 11-29 below, A is a sharply pointing part of a charged conductor C, an object B being placed in front of A. Because of the high surface density of charge at A and the associated electric field of high intensity near A (let us assume for the sake of concreteness that the field is directed from A to B), a current of positively or negatively charged particles suspended in the atmosphere will be set up in the region surrounding A. For positively charged particles, the current will be, according to our assumption above, from A to B while for negatively charged particles it will be from B to A.

Charged particles suspended in the atmosphere are found to exist everywhere. While there occurs a preponderance of positively charged particles at some places, negatively charged particles are found to be more numerous at some others.

Moreover, if the charge density at A is negative (in which case the field is from B to A) and the field intensity happens to be sufficiently large near that point, then electrons can be released at A from the conductor itself, which then join in the current of the charged particles near A. As a result of this current, part of the charge in C gets neutralized while the object B in front of A gets charged. This process of charge *spraying* is often a convenient one for the charging of bodies. For sufficiently intense electric fields set up near a sharply pointed part of a conductor, an *electrical discharge* can also take place, where the atoms in the atmosphere close to A get *ionized*, or excited to relatively high energy states.

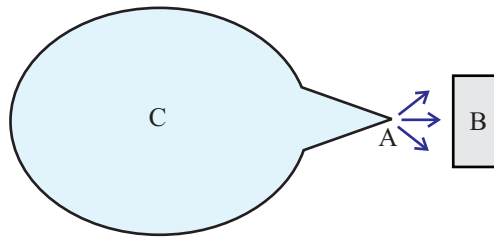


Figure 11-29: An object (B) in front of a sharply pointing part (A) of a conductor (C); an intense electric field is set up near A due to the accumulation of charge in this part of the conductor; assuming that the field is from A to B, a current of positively charged particles is set up in the direction of the arrows (negatively charged particles flow in the opposite direction) resulting in the electrification of B and a neutralization of the charge of C.

11.10.5 Polarization in a dielectric medium

In contrast to a conductor, a dielectric material does not contain free electrons.

If a static electric field is set up in a dielectric medium then no transport of charges takes place in it, but there occurs a relative displacement between the positive and negative charges which, however, continue to remain bound to each other. Consequently, if one looks at any small volume element in the dielectric, one will find that a *dipole moment* has been developed in it. This phenomenon of generation of dipole moments throughout the volume of the dielectric is termed dielectric *polarization*.

In other words, an electric field can be set up in a dielectric medium under electrostatic conditions, when the dielectric becomes *polarized*, with every small volume element in it developing a dipole moment.

The crystalline dielectric materials need a special mention here. In reality, the electrons in a crystalline dielectric cannot be said to be bound to specific atoms or groups of atoms. Because of the regular arrangement of atoms in a crystal, an electron can actually be found everywhere within it with, however, a characteristic probability distribution. The principal difference with an electron in a conducting material lies in the fact that the electrons in a dielectric collectively prevent one another from acquiring energy from an impressed electric field and becoming mobile in the sense of being able to set up a current. As a result, the average velocity or momentum of the electrons in the dielectric continues to be zero in the presence of the electric field.

In an amorphous dielectric all the electrons being bound to specific atoms or groups of atoms, the electrons remain similarly immobile. Consequently, the self-adjusting process referred to above in the case of a conductor whereby the electric field intensity and the charge density become zero, does not occur in a dielectric, either crystalline or amorphous. In other words, a non-zero electric field *can* be set up within a dielectric medium.

11.10.5.1 The polarization vector. Electric susceptibility.

In this context, it may help to look at fig. 11-30 which depicts a region of space R occupied by a dielectric material, and a number of charges (q_1, q_2, q_3 in the figure) that may or may not be placed within the dielectric material itself. These are termed *free* charges since these are not intrinsic to the dielectric material, i.e., are not related to the charges making up the atoms and molecules of the material. If the dielectric material were not there, these free charges would have produced an electric field of intensity, say E_0 within the region R , where E_0 may vary from point to point in R . However, because of the presence of the dielectric material, the actual intensity of the field set up in the region R will be different from E_0 , because R will now contain a distribution of tiny dipoles, with dipole moment, say, P per unit volume around any given point r in the region. P is then referred to as the *polarization* at the point r .

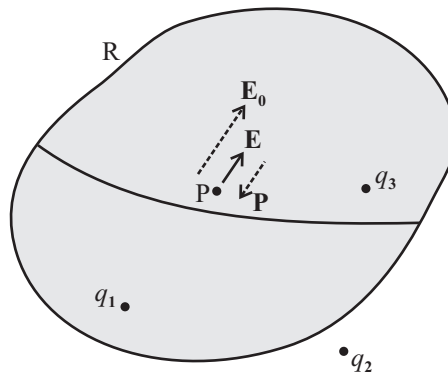


Figure 11-30: Polarization in a dielectric produced by a number of free charges (q_1, q_2, q_3); the electric field E at any point differs from the vacuum field E_0 , due to the polarization in the dielectric; the polarization vector P is, in general opposite to E_0 , cancelling part of the vacuum field so as to produce the resultant field E in the region R occupied by the dielectric.

In the special case when \mathbf{E}_0 is a *uniform* field in a homogeneous dielectric, the polarization \mathbf{P} is also uniform (since the field produced by the free charges affects all the bound charges of the dielectric in an identical manner), and is proportional to \mathbf{E}_0 , and also to the field \mathbf{E} , once again a uniform one, that is actually set up in the region R in the presence of the dielectric. The proportionality constant between the two is expressed in the form $\epsilon_0\chi$, where χ a positive constant characterizing the dielectric material under consideration, referred to as its *electric susceptibility*. In other words,

$$\mathbf{P} = \epsilon_0\chi\mathbf{E}. \quad (11-74)$$

More generally, the relation (11-74) holds even when the field \mathbf{E} is non-uniform, but now it expresses a *local* relation between the polarization vector \mathbf{P} and the electric intensity \mathbf{E} at every point in the dielectric. For a *homogeneous* dielectric, the susceptibility χ is the same at all points in it.

The *linear* relation between the field strength \mathbf{E} and the polarization \mathbf{P} (which can be looked upon as the *response* of the dielectric medium to the field \mathbf{E}) is, in reality, an *approximate* one. In a more complete description, the response \mathbf{P} involves *non-linear* terms in \mathbf{E} that, under ordinary circumstances, are small compared to the linear term and can be ignored as being of little consequence. For *strong* fields, however, such as the field produced by an intense laser light, the non-linear terms become relevant, and make possible a number of applications of great importance.

11.10.5.2 Electric field intensity and the displacement vectors

Referring to the vacuum field \mathbf{E}_0 and the actual field \mathbf{E} in a dielectric, which we assume to be a homogeneous one for the sake of simplicity, the polarization of the medium is found to result in a partial cancellation of the vacuum field \mathbf{E}_0 so that the magnitude of the field (\mathbf{E}) that is actually set up is less than the vacuum field, i.e., $|\mathbf{E}| < |\mathbf{E}_0|$.

Consider, for example, an infinitely extended homogeneous dielectric medium in which a point charge q is placed at the origin O . This constitutes the free charge in the present instance. The field that this free charge would have produced in vacuum at any field

point, say, \mathbf{r} is given by

$$\mathbf{E}_0 = \frac{1}{4\pi\epsilon_0} \frac{q}{r^3} \mathbf{r}, \quad (11-75a)$$

The field that is actually produced on partial cancellation of this vacuum field due to the polarization of the dielectric can be expressed in the form

$$\mathbf{E} = \frac{1}{4\pi\epsilon_0 \epsilon_r} \frac{q}{r^3} \mathbf{r}, \quad (11-75b)$$

where ϵ_r is a constant, termed the *relative permittivity* of the dielectric, given by

$$\epsilon_r = 1 + \chi. \quad (11-76)$$

In other words, the actual field set up in the dielectric is less than the vacuum field by the factor ϵ_r which, by eq. (11-76), is larger than 1 (recall that $\chi > 0$).

The vector $\epsilon_0 \mathbf{E}_0$ in the above example is referred to as the *displacement* vector in the dielectric. Making use of the above relations, one obtains, in the present instance,

$$\mathbf{D} = \epsilon_0 \mathbf{E}_0 = \frac{1}{4\pi} \frac{q}{r^3} \mathbf{r}. \quad (11-77)$$

The relation between the displacement vector and the ‘vacuum field’ is, strictly speaking, not a general one since it holds only for an infinitely extended medium with source charges located within a finite region, where the two satisfy the same *boundary conditions* at infinitely large distances.

The *same* expression is seen to follow on evaluating $\epsilon_0 \mathbf{E} + \mathbf{P}$, by making use of eq. (11-74) and (11-76) (check this out), i.e., the displacement vector (\mathbf{D}) is given by

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P} = \epsilon_0 \epsilon_r \mathbf{E}, \quad (11-78)$$

this formula being of more general validity than the definition in terms of the ‘vacuum

field'. With the displacement vector defined in this manner, the basic principles in the electrostatics of a dielectric can be stated as follows: (a) the electric field intensity \mathbf{E} is a conservative vector field, and (b) the surface integral of the displacement vector over any closed surface equals the total *free* charge residing in the interior of the surface, this being the modified form of Gauss' principle in the case of a dielectric.

Problem 11-10

The electric field intensity at a point $r_1 = 0.6$ m from the center of a uniformly charged spherical body of radius $R = 0.4$ m is $E_1 = 2000$ V·m⁻¹. Calculate the charge density (ρ) of the body, and the potential V at a point at a distance $r_2 = 0.2$ m away from the center. Assume that the body is made of a dielectric material of relative permittivity $\epsilon_r = 3.0$.

Answer to Problem 11-10

HINT: The total charge of the body is $Q = \frac{4}{3}\pi\rho R^3$ and the field intensity at the external point at a distance r_1 from the centre is given by $E_1 = \frac{Q}{4\pi\epsilon_0 r_1^2}$ (see eq. (11-60)). Making use of given values of the relevant quantities, one obtains $\rho = \frac{3\epsilon_0 E_1 r_1^2}{R^3} = 2.99 \times 10^{-7}$ C·m⁻³ (approx). The field intensity at an interior point at distance x , on the other hand, is given by $E_2(x) = \frac{q(x)}{4\pi\epsilon_r\epsilon_0 x^2}$ (see eq. (11-62)) where, for a uniform charge distribution, $q(x) = \frac{4}{3}\pi x^3 \rho$. Making use of formula (11-63) and breaking up the path of integration into a part running from infinity to R and another from R to r_2 , the potential at an internal point at a distance r_2 is seen to be

$$V(r_2) = \frac{Q}{4\pi\epsilon_0 R} - \int_R^{r_2} E_2(x) dx,$$

or,

$$V(r_2) = \frac{Q}{4\pi\epsilon_0 R} \left(1 + \frac{R^2 - r_2^2}{2\epsilon_r R^2}\right) = \frac{E_1 r_1^2}{R} \left(1 + \frac{R^2 - r_2^2}{2\epsilon_r R^2}\right).$$

Using given values of the various quantities, one obtains $V = 2027$ V.

11.10.5.3 Field variables: the question of nomenclature

A few words are in order on the terminology relating to the electrical field variables and, more generally, to electrical and magnetic variables taken together, because of a certain lack of uniformity of usage, especially for the magnetic variables.

In this book, the vector \mathbf{E} representing the electric field strength is mostly referred to as the *electric field intensity* or, in short, the ‘electric intensity’ (at times, further shortened to, simply, the *intensity*). The term ‘electric field strength’ is also in common use. The use of the term ‘intensity’ is in potential conflict with the more common usage of the same term to represent the rate of flow per unit area of field energy, but in reality, the meaning can be unambiguously read from the context. The nomenclature for magnetic field variables will be briefly discussed in sec. 12.8.5.

11.10.5.4 Electric field in a dielectric: summary

These results, stated for the case of a single point charge in a homogeneous dielectric, hold more generally for a distribution of free charges in a homogeneous dielectric. *In summary*, for any given free charge distribution, the field intensity (\mathbf{E}) at any given point in the dielectric differs from the vacuum field (\mathbf{E}_0) that would have been produced if the dielectric were not there, due to the polarization in the dielectric, the latter being quantitatively expressed by the polarization vector (\mathbf{P}), i.e., the dipole moment per unit volume, which is related to \mathbf{E} by eq. (11-74). The material property of the dielectric is expressed in the susceptibility χ and the relative permittivity ϵ_r , the two being related by eq. (11-76).

The electrostatics of dielectrics involves, in addition to the field intensity \mathbf{E} , the specification of one other vector field, namely the displacement vector \mathbf{D} (or, alternatively, the vectors \mathbf{E} and \mathbf{P}). Indeed, for a given distribution of free charges, the field intensity \mathbf{E} in the dielectric is determined by *first* determining the displacement vector and then obtaining \mathbf{E} from eq. (11-78).

In simple situations involving dielectrics, \mathbf{D} is determined by working out the vacuum field \mathbf{E}_0 for the given distribution of free charges and then making use of the relation $\mathbf{D} = \epsilon_0 \mathbf{E}_0$. More generally, however, the determination of \mathbf{D} requires additional considerations involving appropriate *boundary considerations*.

11.10.5.5 Field variables as space- and time averages

The field vectors \mathbf{D} and \mathbf{E} , and the polarization vector \mathbf{P} are all *macroscopically* defined quantities, i.e., ones defined with reference to small volume elements of macroscopic proportions. For instance, if one considers the microscopic charges in a dielectric, it will be found that the charge distribution is characterized by wild fluctuations over distances of inter-atomic separation while, at the same time the charge distribution varies rapidly over extremely small time intervals as well. The electric field intensity resulting from such a fluctuating charge distribution will naturally possess correspondingly sharp fluctuations over microscopic distances in spaces and in extremely small time intervals.

However, commonly used measuring instruments cannot detect or record faithfully such sharp fluctuations. Instead, what the instruments detect is a space- and time *average* of these fluctuations. The principle of superposition ensures that if an average polarization vector is appropriately defined along with an average electric field intensity, and if the displacement vector is defined as in (11-78), then the basic principles of the electrostatics of a dielectric can indeed be expressed in the form of the two statements ((a) and (b)) mentioned at the end of sec. 11.10.5.2.

11.10.5.6 A brief note on relative permittivity

The relative permittivity ϵ_r (also referred to as the dielectric constant) of a medium is determined by its material properties and depends on a large number of factors at the atomic and molecular level. There may be quite a few complexities associated with it that need to be mentioned at this stage. First, the dielectric constant *need not be a scalar* quantity, though it can indeed be considered a scalar for a number of familiar materials. For a number of other materials, however, notably for crystalline ones, the dielectric constant may, more precisely, be described as a *tensor* quantity having a more or less strong directionality associated with it.

What is more, the dielectric constant may not, strictly speaking, be a constant at all, but may depend on the strength of the electric field set up in the material under consid-

eration. A medium with such a field dependence of the dielectric constant is referred to as a *nonlinear* one.

Finally, the dielectric constant pertaining to *static* electric fields in a medium is, truly speaking, a *limiting value* of a more general *function* $\epsilon_r(\omega)$, corresponding to the limit $\omega \rightarrow 0$, where $\epsilon_r(\omega)$ stands for the relative permittivity of the medium pertaining to a time harmonic field of the angular frequency ω that may be set up in the medium.

Such a harmonic or *monochromatic* field (in the form of a monochromatic plane wave, for instance) will be discussed in chapter 14, and will be of primary interest in connection with phenomena relating to the propagation of optical radiation through the medium (chapter 15). A feature closely related to the frequency dependence of the dielectric constant is that ϵ_r need not be a real function of ω , being, more generally, a *complex* function, having an imaginary part. The latter accounts for the *absorption* of electromagnetic waves propagating through the medium.

11.11 Capacitors and capacitance

11.11.1 Charges and potentials on a pair of conductors

Suppose that a pair of conductors A and B have been given charges q_1 and q_2 respectively and that these two, while exerting electrical influence on each other, are far removed from other bodies so as to be free of any other electrical influence, being, at the same time, incapable of sharing their charges with other bodies. The charges q_1 and q_2 will then get distributed on the surfaces of the conductors in such a way that the interior and boundary surface of each conductor are at a constant and uniform potential, say, V_1 and V_2 respectively. Recall that this is the condition for electrostatic equilibrium in respect of charges on conductors.

In accordance with Coulomb's law and the principle of superposition, the relation between the charges q_1 and q_2 and the potentials V_1 and V_2 will be a *linear* one, i.e., of the

form

$$q_1 = c_{11}V_1 + c_{12}V_2, \quad (11-79a)$$

$$q_2 = c_{21}V_1 + c_{22}V_2. \quad (11-79b)$$

In these equations, c_{11} , c_{12} , c_{21} , c_{22} are constants that do not depend on the charges or the potentials on the conductors, and are determined solely by their geometry like their shapes and relative positions. These are termed the *coefficients of capacitance* of the pair of conductors under consideration. For instance, in fig. 11-31 below, two different geometries of arrangement of a given pair of conductors A and B are shown. Though the conductors are the same in the two cases, their coefficients of capacitance will be different.

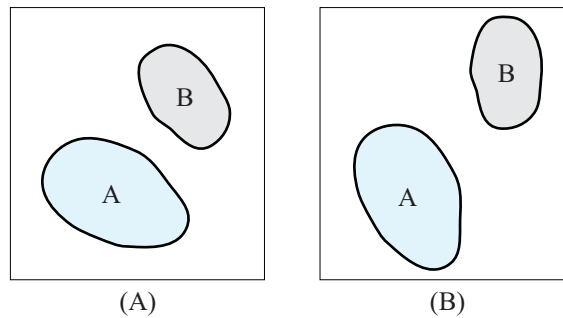


Figure 11-31: Illustrating the concept of coefficients of capacitance; (A) a pair of conductors A and B isolated from the possible influence of other charges and fields; the charges and potentials of these two are related linearly; (B) the same two conductors with a different geometry of arrangement between them; the coefficients relating the charges on these conductors to their potentials are now different.

The situation that has been referred to above is one where the charges q_1 and q_2 on the two conductors are taken as given, it being assumed that the conductors do not share their charges with other bodies. In that case, their potentials V_1 and V_2 are determined uniquely from equations (11-79a) and (11-79b).

One could equally well refer to a different situation where the *potentials* V_1 and V_2

are given by requiring the conductors to be connected to two *sources*. Here the term ‘sources’ is used in the following sense. Imagine two *large* conductors (say, S_1 and S_2) at potentials V_1 and V_2 far removed from the conductors A and B under consideration and connected respectively to these with the help of conducting wires (see fig. 11-32).

In this case the conductors A and B can exchange charges with S_1 and S_2 respectively through the connecting wires, but the potentials of S_1 and S_2 do not change much by this charge exchange because of their large size (which is why we have referred to these as potential *sources*). The connecting wires ensure that A and B acquire the potentials of the sources, namely V_1 and V_2 respectively. Whatever charges were there on A and B prior to the connection will not matter here because there occurs charge exchange through the connecting wires. However, once the electrostatic condition has been reached after establishing the connections, the charges q_1 and q_2 on A and B will be uniquely determined again by equations (11-79a) and (11-79b), with the coefficients c_{11} , c_{12} , c_{21} , c_{22} determined solely by the geometry of the conductors and their relative positions.

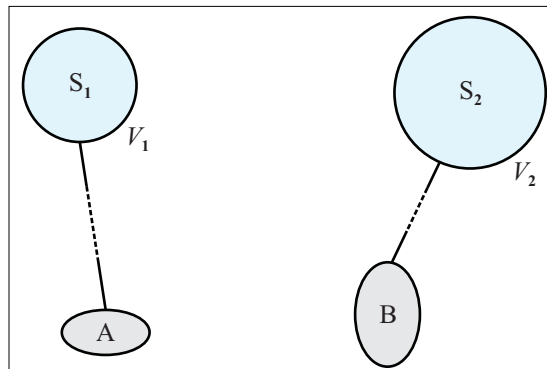


Figure 11-32: Establishing voltages V_1 and V_2 on conductors A and B by connecting them to sources S_1 and S_2 ; the latter are two large conductors far removed from A and B, and connected to these by conducting wires; A and B exchange charges with S_1 and S_2 respectively, while acquiring potentials V_1 and V_2 .

1. At times, the *earth* is used as a convenient potential source. The terrestrial globe can be looked upon, in an approximate sense, as a very large spherical conductor whose potential (say, V_0) is not affected appreciably by charge exchange with other

bodies. The process of establishing electrical contact between various conducting bodies and the earth is referred to as *earthing*.

One can thus say that regardless of the charge on a conductor prior to earthing, its potential after earthing will be V_0 and its charge after earthing will be determined by this potential as also by the possible electrical effect of other bodies.

2. The potential of the earth is often taken to be zero, while the value I have referred to above is V_0 . Whether the potential will be 0 or some other value (say, V_0), will depend on the *reference* point chosen for the potential (see section 11.4.2). As I have mentioned in section 11.4.4, it is often convenient to take this reference point to be a point at infinity. However, in that case one can no longer assume the potential of the earth to be zero. Indeed, the potential of the earth with respect to a point at infinity is *not* zero or of negligible value. How, then, can one take the potential of the earth to be zero? I will briefly address this question in sec. 11.12.
3. In order to ensure that a conductor acquire a desired potential, it is not always necessary to connect it to a distant conductor of large size. Suppose that the negative terminal of an electrical *cell* has been connected to the earth, while the other terminal is connected to the conductor under consideration (see chapter 12 for an introduction to electrical cells). Then, if the electromotive force of the cell be E , the potential of the conductor will be $V_0 + E$.

It can thus be said that, two of the four quantities q_1 , q_2 , V_1 , and V_2 , can be given desired values by appropriate arrangements, when the other two will be determined uniquely in accordance with equations (11-79a) and (11-79b) (in this context, see sec. 11.11.2 below). The two independently fixed quantities may be, for instance, q_1 and q_2 or, say, V_1 and V_2 . Or again, these may even be, say, q_1 and V_2 or q_2 and V_1 . In each case, however, equations (11-79a) and (11-79b) determine the remaining two of the above four quantities. As I have mentioned, the four coefficients of capacitance occurring in the above relations depend only on the geometry of the two conductors involved. However, not all of these four quantities are independent. In reality, only three of the four coefficients can be made to vary independently since considerations relating to the energy of the system made up of the pair of charged conductors show that c_{12} and c_{21} have to be equal to each other.

The above considerations can be extended to include situations involving more than two conductors. Thus, considering, N number of conductors C_1, C_2, \dots, C_N , there obtains a relation of the form

$$q_i = \sum_{j=1}^N c_{ij} V_j \quad (i = 1, 2 \dots N), \quad (11-80)$$

where the V 's stand for the potentials of the conductors, the q 's for their charges, and the c 's for the relevant coefficients of capacitance, which satisfy the symmetry relations $c_{ij} = c_{ji}$ ($i, j = 1, 2, \dots N$). The coefficients of capacitance are determined solely by the geometries of the conductors and their dispositions in space.

11.11.2 The Uniqueness Theorem

In this context the *uniqueness theorem* of electrostatics may be taken note of. One version of the theorem states that if a region R in space is free of charges, and if the potential at every point of the boundary surface of R (the boundary surface may be made up several disjoint components, depending on the geometry of the region R) is specified, then the field intensity everywhere in R is uniquely determined. Thus, considering a number of conductors at given potentials, and taking R to be the region outside the volumes occupied by the conductors (see figure 11-33), one can apply the theorem by noting that the boundary surface of R is made up of the outer surfaces of the conductors and a 'surface at infinity', the latter being a closed surface, every point of which is at an infinitely large distance from each of the conductors (the conductors themselves are assumed to be confined within a finite region of space).

Since the potential at every point on the surface at infinity is zero and that on each of the surfaces of the conductors is also specified, the only additional specification that one needs for the uniqueness theorem to apply is that the region R be free of source charges. Assuming, then, the absence of source charges in R , the theorem implies that the electric field intensity is uniquely determined everywhere which also implies, in particular, that the normal component of the field vector at every point on the surfaces of the conductors is also uniquely determined (the tangential component is zero since

the conductors are equipotential surfaces). This, in turn, means that the charge density at every such point is uniquely determined, implying uniquely determined values of the total charges on the conductors.

By virtue of the uniqueness theorem, The coefficients of capacitance for the given set of conductors, introduced in sec. 11.11.1 are uniquely determined by the geometries of the conductors and their dispositions in space.

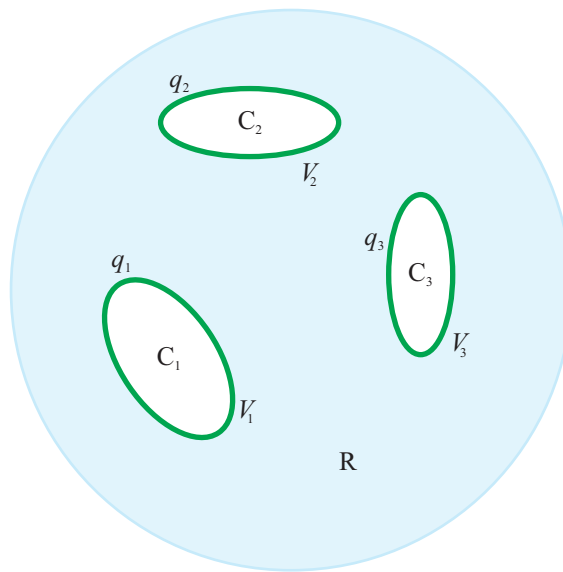


Figure 11-33: Explaining the idea of the uniqueness theorem as applying to conductors held at specified potentials; C_1, C_2, C_3 are conductors with specified dispositions (the theorem applies to an arbitrary number of conductors), located within a finite region of space; the region R exterior to the conductors is free of source charges; if the potentials (V_1, V_2, V_3) of the conductors are specified, then the electric field strength is uniquely determined everywhere in R ; this implies, in particular, that the normal component of the field is determined uniquely at each and every point on the boundary surfaces of the conductors; thus, the charge distributions on the conductor surfaces and the total charges (q_1, q_2, q_3) are also uniquely determined; the boundary of the region R can be looked upon as being made up of the boundary surfaces of the conductors and a 'surface at infinity' on which the potential is zero.

Digression: Electrostatic shielding

An interesting consequence of the uniqueness theorem, applied to a set of hollow conductors with a given disposition in space can now be stated. Fig. 11-34 depicts a hollow conductor C , where the regions exterior and interior to C are marked R_1, R_2 re-

spectively. Let the interior region (R_2) be free of source charges. Suppose that the outer surface (S) of the conductor is at a specified potential V , caused by any one or more of the following: (i) source charges placed in the region R_1 , (ii) charge given to the conductor C , and (iii) a voltage source connected to C . Then, so long as the potential V remains fixed, the field in the interior region R_2 will be zero, with the potential on the interior surface S' of the conductor remaining at the value V . In other words, the interior of the conductor is *shielded* from whatever rearrangements are made in the source charges exterior to C , so long as V remains unchanged.

As a related instance of shielding, consider the situation where one or more source charges, which we assume to be point charges for the sake of concreteness, are placed in the interior region R_2 of the hollow conductor C , while the exterior region R_1 is free of source charges. If the total charge in R_2 be Q then the potential V of the conductor and the field in the exterior region R_1 is the same as if there were no charges in R_2 , and a charge Q were given to C (this charge gets distributed on the exterior boundary (S) of C , the surface density being determined by the geometry of S). In other words, the potential of the conductor, the surface distribution of charges on S and the field in the exterior region R_1 are independent of (i.e., are shielded from) the disposition of the charges in the interior region R_2 , being solely determined by the *total* charge Q . The field in the interior region and the induced charge distribution over the interior boundary (S'), however, will depend on the disposition of the interior charges and the geometry of S' (the total induced charge on S' will be $-Q$).

This shielding effect of conductors, where interior and exterior regions are shielded from electrostatic effects of the exterior and the interior regions respectively, holds for more than one conductors as well.

11.11.3 Capacitance of a pair of conductors

A pair of conductors insulated from each other and isolated from the electrical effect of other bodies is referred to as a *capacitor*, while the term *condenser* is also used. Imagine a situation in which the charges on the conductors are equal and opposite (say,

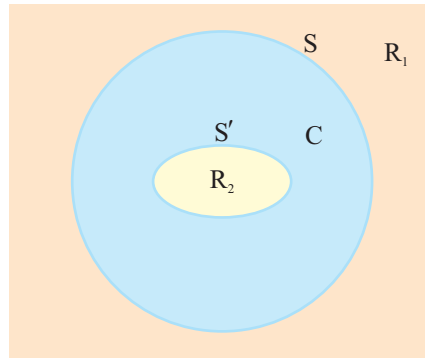


Figure 11-34: Explaining the idea of electrostatic shielding; C is a hollow conductor with exterior and interior boundary surfaces S, S'; R_1 and R_2 are the regions exterior and interior to C; if there are no source charges in R_2 and if the potential V of C is specified, then the field within R_2 is zero, regardless of source charges in R_1 ; the charge distribution on S depends on its geometry, but that on S' is zero; likewise, if source charges are placed within R_2 and there are no source charges in R_1 , then the potential of C and the field in R_1 are determined only by the total charge in R_2 , and not by the distribution of the charges; in this sense, the interior and exterior regions are shielded from each other

$q_1 = q$ and $q_2 = -q$). If now V denotes the potential difference between the conductors ($V = V_1 - V_2$), then the ratio

$$C = \frac{q}{V}, \quad (11-81)$$

is referred to as the *capacitance* of the capacitor. Like the coefficients of capacitance, its value is once again determined solely by the geometry of the conductors. Indeed, as may be expected, the capacitance is related to the coefficients of capacitance of the two conductors:

$$C = \frac{c_{11}c_{22} - c_{12}^2}{c_{11} + c_{22} + 2C_{12}}. \quad (11-82)$$

Problem 11-11

Check formula (11-82) out.

Answer to Problem 11-11

HINT: Defining $V' = \frac{V_1 + V_2}{2}$, and substituting $q_1 = q$, $q_2 = -q$, $V_1 - V_2 = V$ in (11-79a), (11-79b),

one gets

$$q = c_{11}(V' + \frac{V}{2}) + c_{12}(V' - \frac{V}{2}) = (c_{11} + c_{12})V' + \frac{1}{2}(c_{11} - c_{12})V, \quad (11-83a)$$

$$-q = c_{21}(V' + \frac{V}{2}) + c_{22}(V' - \frac{V}{2}) = (c_{21} + c_{22})V' + \frac{1}{2}(c_{21} - c_{22})V. \quad (11-83b)$$

Elimination of V' gives (11-82) for $C = \frac{q}{V}$.

The unit of capacitance is $C \cdot V^{-1}$, also termed the *farad* (F). For practical purposes, the μF (microfarad, 10^{-6}F), nF (nanofarad, 10^{-9}F), and the pF (picofarad, 10^{-12}F) are, at times, more convenient to use.

11.11.4 The spherical condenser

As an example of a capacitor and its capacitance, I refer first to a *spherical condenser*. Suppose that an insulated conducting sphere of radius a is given a charge q and is removed away from the electrical influence of other bodies. What will then be its potential?

The answer to this question is not difficult to arrive at, once one applies Gauss' principle to the problem. The result one arrives at is (see sec. 11.9.1)

$$V = \frac{q}{4\pi\epsilon_0 a}. \quad (11-84a)$$

One then says that the *capacitance* of the spherical condenser is

$$C = \frac{q}{V} = 4\pi\epsilon_0 a. \quad (11-84b)$$

However, the question that comes up here is, how does the definition of a capacitor and its capacitance apply for one single conductor when the definition given above refers to a *pair* of conductors to start with? In order to answer this, I address first the problem of a pair of concentric spherical conductors.

Problem 11-12

An insulated spherical conductor of radius $a = 0.1\text{m}$ is given a charge $q = 1.5 \times 10^{-7}\text{C}$. calculate the force per unit area due to the field created by the conductor, at any point chosen point on it.

Answer to Problem 11-12

HINT: Due the spherical symmetry, the charge q is uniformly distributed over the surface of the conductor, which implies that the surface charge density is $\sigma = \frac{q}{4\pi a^2}$. Hence, the surface density of force at any point on the conductor, acting radially outward, is, by formula (11-73), $f = \frac{q^2}{32\epsilon_0\pi^2 a^4}$. Substituting given values, one obtains $f = 0.08 \text{ N}\cdot\text{m}^{-2}$.

11.11.5 A pair of concentric spherical conductors

Fig. 11-35 depicts a pair of conductors, insulated from each other, in the form of concentric spherical shells with the inner conductor having inner and outer radii a' and a respectively, and the outer conductor having corresponding radii b and b' ($a' < a < b < b'$). Let charges q_1 and q_2 be given to the two conductors, and let their potentials be V_1 and V_2 .

The charge q_1 will spread itself uniformly over the outer surface of the inner conductor C_1 , while the charges on the inner and the outer surfaces of the outer conductor C_2 will be $-q_1$ and $q_1 + q_2$ respectively.

1. Here I skip the proofs of these statements. The fact that the charges on the outer surface of C_1 and the inner surface of C_2 are equal and opposite will be seen to follow from considerations below.
2. Supposing that a charge q' is placed in the hollow of the inner conductor C_1 , and a charge q'' has been given to the body of C_1 , the charge on its outer surface will be $q' + q''$ distributed uniformly, in which case q_1 will have to be taken as $q' + q''$.

In accordance with Gauss' principle, the field intensity at any point \mathbf{r} in the region between the two conductors (i.e., for $a < r < b$; here \mathbf{r} stands for the position vector with respect to the center O of the spheres chosen as the origin) is $\mathbf{E} = \frac{q_1}{4\pi\epsilon_0 r^3}\mathbf{r}$, regardless

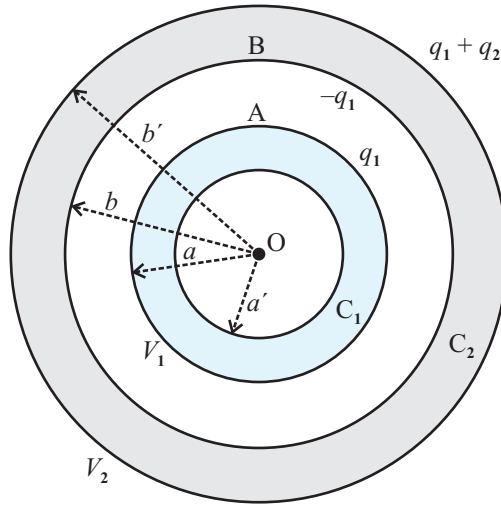


Figure 11-35: A pair of conductors in the form of concentric spherical shells; the inner and outer radii of the inner shell are a' and a , while the corresponding radii for the outer conductor are b and b' ; charges q_1 and q_2 are given to the conductors, which get distributed on the inner and outer surfaces as shown (we assume for the sake of simplicity that there is no charge residing in the hollow of the inner conductor); the potentials of the conductors are V_1 and V_2 ; the coefficients of capacitance between the conductors, as also their capacitance can be determined by working out the relation between q_1 , q_2 , and V_1 , V_2 , and are given by (11-87), (11-86).

of the charges on the inner and outer surfaces of C_2 (refer to eq. (11-57)). On taking the line integral of the field along a radial line joining two points A and B shown in the figure, one finds

$$V_1 - V_2 = \frac{q_1}{4\pi\epsilon_0} \frac{b - a}{ab}, \quad (11-85)$$

(check this out).

In particular, if the charges given to the two conductors be q and $-q$ respectively, then the charge on the outer surface of C_2 will be zero and the potential difference between the two conductors will be given by eq. (11-85), with q_1 replaced with q . The capacitance of the conductors will then be, by definition

$$C = \frac{4\pi\epsilon_0 ab}{b - a}. \quad (11-86)$$

Referring to the more general situation in which the charges on the conductors are q_1 and q_2 , one can invoke Gauss' principle in several steps to work out the potentials V_1

and V_2 , from which the coefficients of capacitance of the two conductors can be worked out. I quote the results here:

$$c_{11} = \frac{4\pi\epsilon_0}{\frac{1}{a} - \frac{1}{b}}, \quad c_{22} = b' \left(\frac{1}{a} - \frac{1}{b} + \frac{1}{b'} \right) c_{11}, \quad c_{12} = c_{21} = -c_{11}. \quad (11-87)$$

With these expressions for the coefficients of capacitance one can check that the capacitance, as determined from eq. (11-82) agrees with that in eq. (11-86).

In this case, the potentials of the inner and outer conductors are given by

$$V_1 = \frac{1}{4\pi\epsilon_0} \left(q_1 \left(\frac{1}{a} - \frac{1}{b} + \frac{1}{b'} \right) + q_2 \left(\frac{1}{b'} \right) \right), \quad (11-88a)$$

$$V_2 = \frac{1}{4\pi\epsilon_0} \frac{q_1 + q_2}{b'}. \quad (11-88b)$$

All the results derived in this section are with reference to a pair of spherical conductors where the region in between the two spherical surfaces is vacuum, i.e., devoid of any material medium. The effect of a dielectric medium occupying the region between the two conductors will be considered in sec. 11.11.12.

Problem 11-13

Establish the relations in (11-87) and show that the formula for capacitance, eq. (11-86) follows.

Answer to Problem 11-13

HINT: The charges on the two spheres constitute a spherically symmetric distribution, and hence the potential on the outer surface of the outer conductor B is, by formula (11-64a), $V_2 = \frac{q_1 + q_2}{4\pi\epsilon_0 b'}$ (taking $r \rightarrow b'$; this verifies formula (11-88b)). Since a conductor is an equipotential body, this must also be the potential at $r = b$. By Gauss' theorem, the field in region $a < r < b$ is $E(r) = \frac{q_1}{4\pi\epsilon_0 r^2}$, directed radially, and hence the potential at $r = a$ is $V(a') = V_1 = V_2 - \int_b^a \frac{q_1}{4\pi\epsilon_0 r^2} dr$, which verifies (11-88a). We now invert formulae (11-88a), (11-88b), to obtain q_1, q_2 as linear combinations of V_1, V_2 , and compare with (11-79a), (11-79b) to obtain the relations (11-87). Substituting in formula (11-82), the capacitance of the pair of concentric conducting shells is obtained as

$$C = \frac{c_{11}c_{22} - c_{12}^2}{c_{11} + c_{22} - 2c_{12}} = c_{11}, \text{ verifying (11-86).}$$

11.11.6 Capacitance of a single conductor

Note that, if the inner radius of the outer sphere be large compared to the outer radius of the inner one ($\frac{a}{b} \rightarrow 0$), the capacitance goes to $C = 4\pi\epsilon_0 a$, i.e., the same as the expression in eq. (11-84b).

Indeed, for ($b \rightarrow \infty$), eq. (11-88b) implies that $V_2 \rightarrow 0$ (recall that $b' > b$) and eq. (11-85) reduces to eq. (11-84a). In other words, under this condition, the pair of concentric spherical conductors becomes effectively similar to the single spherical conductor of radius a . Recall that, in deriving eq. (11-84a), the potential at infinity has been assumed to be zero. The expression (11-84b) can then be referred to as a capacitance in the sense that it represents the capacitance of a spherical conductor of radius a , considered with another concentric spherical conductor of infinitely large radius. It is in this sense that we will use the terms capacitor and capacitance for a single spherical conductor.

Similar considerations apply to a single conductor of any other shape. If the conductor, isolated from other electrical effects, be given a charge Q , and if its potential be V , then its capacitance is defined, once again, as $C = \frac{Q}{V}$.

Here, as in the case of a single spherical conductor, one can imagine a hollow conductor surrounding the conductor (C) under consideration (fig. 11-36) where the boundary of this hollow conductor is located at an infinitely large distance (this is indicated in fig. 11-36 by drawing the boundary with a dotted line and using outward arrows; the shape of the boundary is not important here; the distinction between the outer and inner boundaries of this hollow conductor is also not relevant). The capacitance of *this* pair of conductors will then be the same as the capacitance of the single conductor as defined above.

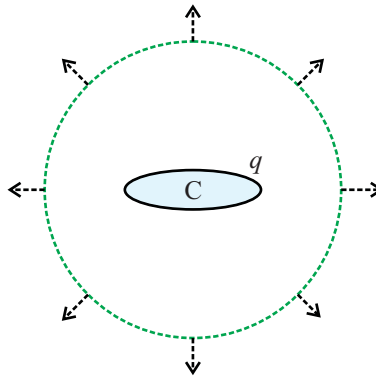


Figure 11-36: A conductor C with a charge q and an imagined conductor surrounding it, with its boundary at an infinitely large distance; the capacitance of this pair can be taken as the definition of the capacitance of the conductor C in isolation; the dotted arrows indicate that the boundary of the imagined outer conductor is to be drawn out to infinity.

11.11.7 Self-capacitance and mutual capacitance

Considering a system made of two conductors (say, C_1 , C_2) insulated from one another and isolated from all other bodies that may possibly have an influence on the charges and potentials of these two, we have seen that this system is characterized by three coefficients of capacitance (c_{11} , c_{22} , and $c_{12} = c_{21}$), where the capacitance C of the pair of conductors is given by eq. (11-82). In addition, each of the conductors, considered in isolation, has a capacitance of its own, as explained in sec. 11.11.6. The two capacitances, say, C_1 and C_2 are termed the self-capacitances of the two conductors. The coefficient c_{12} , on the other hand, on which the capacitance C of the two conductors depends, is referred to as the *mutual capacitance* of the two.

A situation of some interest is one where the coefficient c_{12} is small compared to c_{11} and c_{22} , which happens when the two conductors are separated from each other by a large distance so that a charge given to any one of the conductors affects the potential of that conductor to a much greater extent compared to the potential of the other conductor. In this case, equations (11-79a), (11-79b) imply $q_1 = c_{11}V_1$ and $q_2 = c_{22}V_2$, i.e., the self capacitances C_1 and C_2 are the same as the coefficients c_{11} and c_{22} respectively. Moreover, the capacitance C of the two conductors in this case is given by $C = \frac{C_1 C_2}{C_1 + C_2}$. This can be interpreted by saying that the capacitor made up of the two conductors is equivalent to the *series combination* (see sec. 11.11.11) of the conductors considered

individually as capacitors (in accordance with the considerations in sec. 11.11.6).

The opposite situation corresponds to the case where both the coefficients c_{11} and c_{22} are small compared to the magnitude of the mutual capacitance, $|c_{12}|$. In this case the capacitance C of the conductors is given by $C \approx -\frac{c_{12}}{2}$. This corresponds to a situation where a charge given to C_1 produces only a small change in the potential of C_1 but a considerably larger change in the potential of C_2 , and similarly, a charge given to C_2 affects the potential of C_1 to a considerably greater extent compared to the change in the potential of C_2 itself.

The *parallel plate capacitor* constitutes another interesting special case (see sections 11.11.8, 11.11.8.1 below).

11.11.8 The parallel plate capacitor

Fig. 11-37(A) depicts a parallel plate capacitor where two conducting plates C_1 , C_2 , insulated from each other, are kept at a small distance d apart, the area (A) of either plate being large in the sense of its linear dimension (say, L ($\sim \sqrt{A}$)) being large in comparison to d . If charges q and $-q$ are given to the two plates (fig. 11-37(A)) then these charges get distributed over the *inner surfaces* of the two plates (i.e., the surfaces facing each other), where the charge distribution is, in a certain approximate sense, uniform. The electric field intensity E anywhere between the plates is along the unit vector \hat{n} directed normally from the upper to the lower plate as in the figure, once again in a certain approximate sense, since there occur small deviations from this uniform field, especially near the rim of the plates. The approximate description in terms of a uniform surface charge density (say, σ) and a uniform electric field (along with zero charge on the outer faces of the two conducting plates) becomes exact only in the limit of $\frac{L}{d} \rightarrow \infty$.

Referring to the result (11-72), the field intensity E is given in terms of the surface

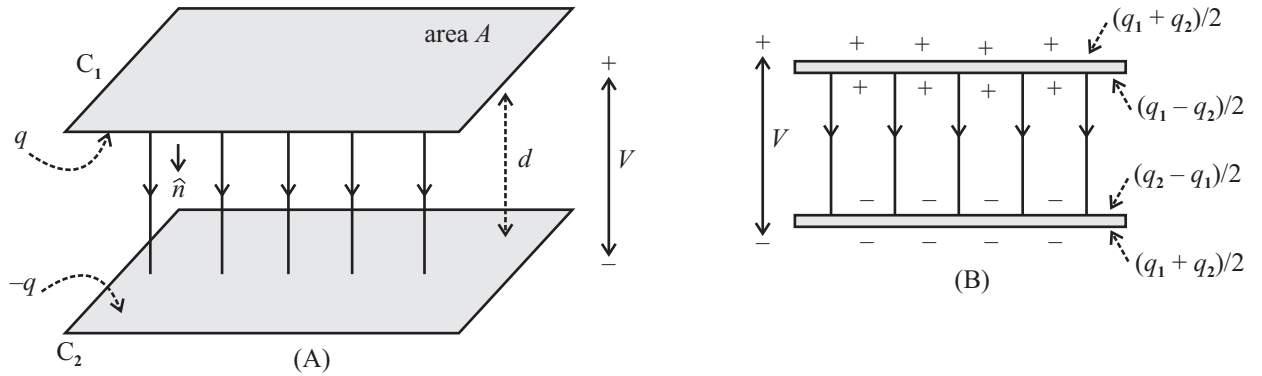


Figure 11-37: (A) The parallel plate capacitor; a number of lines of force between the plates are shown; the charges given to the two plates are q and $-q$, which reside on the inner surfaces of the plates, if the separation d be small compared to \sqrt{A} (where A is the area of either plate); the charge distribution is approximately uniform on the inner surface of either plate; (B) charge distribution between the inner and outer surfaces of the parallel plate condenser, when charges q_1, q_2 are given to the plates; in this more general case, the potentials of the plates are $V_0 + \frac{V}{2}$, $V_0 - \frac{V}{2}$, where the mean potential V_0 is large compared to the potential difference V .

charge density σ (and consequently in terms of q and A) by

$$E = \frac{\sigma}{\epsilon_0} = \frac{q}{\epsilon_0 A}. \quad (11-89)$$

The electric field intensity being the rate of variation of the potential with distance, taken with a negative sign, the potential decreases at a uniform rate from the upper to the lower plate along a direction normal to the plates, and hence the potential difference between the upper and the lower plates is obtained from

$$E = \frac{V}{d}. \quad (11-90)$$

Combining the above two results, the *capacitance* of the parallel plate condenser works out to

$$C = \frac{q}{V} = \frac{\epsilon_0 A}{d}, \quad (11-91)$$

where we have assumed the region between the two plates to be free space, i.e., devoid of any material medium. The effect of a dielectric medium in the region between the two plates will be considered in sec. 11.11.12.

11.11.8.1 Charge distribution in the plates

The parallel plate capacitor is often used with one of the two conducting plates (say, the lower plate in fig. 11-37(A)) earthed, and with a charge, say, q given to the other plate (recall that, of the four quantities q_1 , q_2 , V_1 , V_2 , two can be given chosen values while the remaining two are then determined by the coefficients of capacitance). In this case, only a negligible part of the charge given to the upper plate C_1 resides on its upper surface while almost the entire charge resides on inner surface, i.e., the surface facing C_2 (and an equal and opposite charge is induced on the inner surface of C_2 (i.e., the surface facing C_1)). The upper conductor then acquires a potential $\approx \frac{q}{C}$ relative to the potential of the lower conductor, with C given by (11-91).

The earth potential is commonly assumed to be zero. In reality, however, the potential of the earth has a non-zero value (say, V_0) in which case all other potentials are to be considered with respect to this potential of the earth (see sec. 11.12 for the condition of validity of this statement). Because of a non-zero (and considerably large) earth potential, there will be some charge on the outer faces of the two conducting plates, but the potential of the upper plate with respect to the earth potential will still be given by $V = \frac{q'}{C}$ where q' (differing from q) stands for the charge in the inner surface of the upper conductor C_1 . Correspondingly, the charge induced in the inner surface of the lower plate C_2 will be $-q'$, while the charge on the outer (lower) face of C_2 will be determined by the charged density at the earth's surface, since the field strength in the region between this lower face of C_2 and the earth's surface has to be zero. The charge $(q - q')$ residing on the outer (upper) surface of the upper conductor C_1 turns out to be of the order of $V_0 C' + q \frac{C'}{C}$, where C' is the self-capacitance of the upper plate, and is small compared to C (see below). For practical purposes, one can take $q' \approx q$.

Consider now a situation in which the two plates, maintained at a large distance from the earth's surface, are insulated and are given charges q_1 and q_2 . Generally speaking, some charge will remain on the outer surfaces of the two conducting plates. Approximate expression for these charges as also for the potentials of the two plates can be obtained under the condition that the dimension L ($\sim \sqrt{A}$) is large compared to the

separation (d) between the plates, in which case the coefficients c_{11} , c_{22} and c_{12} can be expressed in terms of the capacitance C of the parallel plate capacitor, and the self-capacitance, say, C' , of either of the two plates considered in isolation from the other, where dimensional arguments indicate that $C' \sim \epsilon_0 L \ll C$ (recall the condition $d \ll \sqrt{A}$ we started with).

The problem of determining the capacitance of a thin plate of arbitrary shape has no known solution, while an approximate expression can be worked out for a thin circular disk, though the derivation is quite non-trivial. One can, for instance, work out the capacitance of a conductor having the shape of an oblate spheroid, the latter being characterized by two semi-axial lengths a , b , where a is the radius of the circular section of the spheroid and b ($< a$) is the minor axis of the elliptical section. The capacitance of the thin circular disk is then obtained by taking the limit $\frac{b}{a} \rightarrow 0$. The result of this calculation turns out to be $8\epsilon_0 a$, up to a correction term due to the small thickness of the disk.

One then finds that, if the potentials (V_1 , V_2) of the two plates be expressed in the form

$$V_1 = V_0 + \frac{V}{2}, \quad V_2 = V_0 - \frac{V}{2}, \quad (11-92)$$

then V_0 , the mean potential of the plates, is *large* compared to V , their potential difference. Indeed, under the approximation mentioned above, the charges residing on the outer surfaces of the two plates can be seen to be $\frac{q_1+q_2}{2}$ each and the charges on the inner surfaces are then $\frac{q_1-q_2}{2}$, $\frac{q_2-q_1}{2}$, as shown in fig. 11-37(B). An approximate expression for V_0 is

$$V_0 \approx \frac{q_1 + q_2}{C'}, \quad (11-93)$$

while V is given by

$$V = \frac{q_1 - q_2}{2C}. \quad (11-94)$$

One then finds that the coefficients of capacitance are given by the approximate expres-

sions

$$c_{11} = c_{22} \approx C + \frac{C'}{4}, \quad c_{12} \approx -C + \frac{C'}{4}, \quad (11-95)$$

where one finds that the relation (11-82) is conformed to. Knowing these approximate expressions for the coefficients of capacitance, and any two of the four quantities q_1 , q_2 , V_1 , V_2 , one can determine approximate expressions for the remaining two under conditions stated above.

1. I will now outline to you as to how the above relations are obtained, starting from the assumption that the self-capacitance C' of either of the two conducting plates is small compared to the capacitance C of the parallel plate capacitor. Let the charge in the upper (outer) surface of the upper plate C_1 be q' . Consider the field in the region above the upper plate, i.e., the region extending from its upper surface to an infinite distance. This field is shielded from the field in the region between the two conducting plates, and is therefore identical to the field produced by a circular disk having a charge $2q'$ (note the factor of 2 which arises because, for an isolated plate, the charge given to it gets divided equally between its two surfaces), and hence one obtains $\frac{2q'}{V_1} = C'$. Similarly, $\frac{2q''}{V_2} = C'$, where q'' is the charge residing on the outer (lower) surface of the lower conducting plate C_2 . The charges on the inner surfaces of the two plates being, respectively $q_1 - q'$ and $q_2 - q''$, we have two more relations of the form $q_2 - q'' = -(q_1 - q')$ (uniformity of the field between the two plates), and $C = \frac{q_1 - q'}{V_1 - V_2}$. Together, these relations give all the results stated above, as you can check for yourself.
2. Incidentally, the charge distribution between the outer and inner surfaces of the two plates given above, is consistent with the requirement that the field strengths *within* each of the two conductors is to be zero (check this out by making use of the result that the field strength on either side of a thin layer of charge of surface density σ at any point at a small distance from the layer is $\frac{\sigma}{2\epsilon_0} \hat{n}$ where \hat{n} is a unit normal to the surface of the layer, directed away from it).

Problem 11-14

Insulated parallel plates A,B,C of identical shape and size are arranged as in fig 11-38, the linear dimension (L) of each being large compared to the separations d_1, d_2 , and charges Q_1, Q_2, Q_3 are given to the plates. Find the charges on the inner and outer surfaces ($S_1, S'_1, S_2, S'_2, S_3, S'_3$) of the plates, and their potentials.

Answer to Problem 11-14

HINT: The self-capacitance (C) of each plate, considered in isolation, enters into the problem in an essential way. Under the given conditions, C is small compared to the capacitances $C_1 = \frac{\epsilon_0 A}{d_1}, C_2 = \frac{\epsilon_0 A}{d_2}$ (A = area of each plate) formed by A, B and B, C, and the regions above A and below C are shielded from the regions between A, B, and B,C. Hence the outer surfaces of A, C are equivalent, and the total charge ($Q_1 + Q_2 + Q_3$) gets equally distributed between these surfaces (up to small corrections, see below). With charge $q_1 \approx \frac{Q_1+Q_2+Q_3}{2}$ on the outer surface (S_1) of A, the charge on its inner surface (S'_1) is $q'_1 = Q_1 - q_1$. This induces an equal and opposite charge on S_2 (the field strength between S'_1 and S_2 is uniform), and thus the charges on S_2, S'_2 are $q_2 = -q'_1 = q_1 - Q_1$ and $q'_2 = Q_2 - q_2$. Continuing the reasoning, the charge on S_3, S'_3 are $q_3 = -q'_2 = q_2 - Q_2$ and $q'_3 = Q_3 - q_3$ (reason this out).

The above results are correct to terms of the order of $\frac{C}{C_1}, \frac{C}{C_2}$. The corrections can be worked out as follows. Considering any of the three plates in isolation, a charge q given to it raises it to a potential of $\frac{q}{C}$. Noting that this charge gets divided equally between the two surfaces of the plate, the potential of A has to be $V_1 = \frac{2q_1}{C}$ (reason out why; here q_1 , the charge on S_1 , differs from $\frac{q_1+q_2+q_3}{2}$ by a small correction to be determined), and similarly the potential of C will be $V_3 = \frac{2q'_3}{C}$, where q'_3 , the charge on S'_3 again differs from $\frac{q_1+q_2+q_3}{2}$ by a small correction. But we must have $V_1 = V_3 + \frac{q'_1}{C_1} + \frac{q'_2}{C_2}$ (reason out why). This, along with the relation $q_1 + q'_3 = Q_1 + Q_2 + Q_3$, determines the charges (including correction terms) on all the surfaces. One obtains

$$q_1 = \left(2 + \frac{C}{2C_1} + \frac{C}{2C_2}\right)^{-1} \left[Q_1 \left(1 + \frac{C}{2C_1} + \frac{C}{2C_2}\right) + Q_2 + Q_3 \left(1 + \frac{C}{2C_2}\right)\right],$$

(check this out). With this corrected value of q_1 , the charges on the other surfaces are

$$q'_1 = Q_1 - q_1, \quad q_2 = q_1 - Q_1, \quad q'_2 = Q_1 + Q_2 - q_1,$$

$$q_3 = q_1 - Q_1 - Q_2, \quad q'_3 = Q_1 + Q_2 + Q_3 - q_1.$$

The potentials to which the plates get raised are high. Up to correction terms, all the three

plates are at potential $\frac{Q_1+Q_2+Q_3}{C}$. However, the three potentials differ when correction terms are included. With the corrected value of q_1 and of the other charges obtained above, one now has $V_1 = \frac{2q_1}{C}$, $V_2 = V_1 - \frac{q'_1}{C_1}$, $V_3 = V_2 - \frac{q'_2}{C_2}$.

We assume the plates to be far removed from earth's surface; otherwise the earth's potential will have to be added to the values of V_1, V_2, V_3 obtained above (refer to sec. 11.12 below).

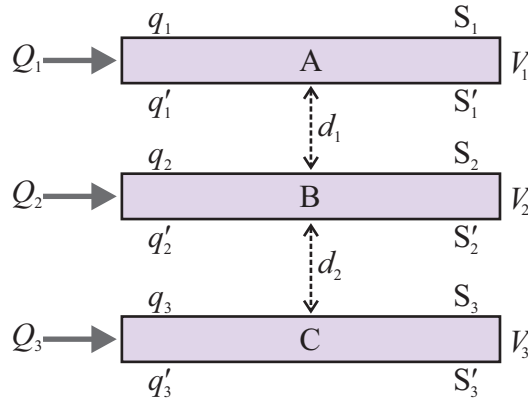


Figure 11-38: Insulated parallel plates A, B, C arranged as specified in problem 11-14, with charges Q_1, Q_2, Q_3 given to the plates; the self-capacitance (C) of each plate enters into the problem in an essential way; under the given conditions, C is small compared to the capacitances C_1, C_2 formed by A, B and B, C, and the regions above A and below C are shielded from the regions between A, B, and B, C; the total charge ($Q_1 + Q_2 + Q_3$) gets equally distributed between the outer surfaces of plates A, C (up to small correction terms); the plates acquire high potentials V_1, V_2, V_3 , while the potential differences $V_1 - V_2, V_2 - V_3$ are relatively small, determined by C_1, C_2 .

11.11.9 Cylindrical capacitor

Fig. 11-39 depicts a pair of long coaxial cylindrical conductors with radii a, b (for hollow cylinders, a is the outer radius of the inner cylinder, while b denotes the inner radius of the outer cylinder). For sufficiently long cylinders, one may take the length to be effectively infinity, in which case, if the inner cylinder is given a charge λ per unit length (we assume λ to be positive for the sake of concreteness), then the field strength at any point in the region between the two conductors is given by formula (11-67) where the field is directed radially away from the inner cylinder (the inner surface of the outer cylinder will acquire an induced charge $-\lambda$ per unit length; this can be seen by evaluating the field strength at $r = b$ and then invoking formula (11-72)). We now integrate from $r = b$

to $r = a$ so as to obtain

$$V_1 - V_2 = -\frac{\lambda}{2\pi\epsilon_0} \int_b^a \frac{dr}{r} = \frac{\lambda}{2\pi\epsilon_0} \ln \frac{b}{a}, \quad (11-96)$$

which tells us that the *capacitance per unit length* of the cylindrical capacitor is given by

$$C = \frac{2\pi\epsilon_0}{\ln \frac{b}{a}}. \quad (11-97)$$

Here, as in earlier sections, we have assumed the region between the two cylindrical conductors to be devoid of any material medium, for the sake of simplicity.

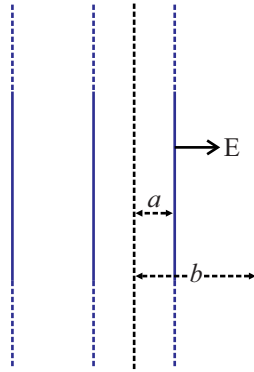


Figure 11-39: The cylindrical capacitor, made up of a pair of long coaxial cylindrical conductors of radii a , b (section by the plane of the figure); only a part of the length of each cylinder is shown, the rest being depicted with dotted lines; the capacitance per unit length is given by formula (11-97); the region between the two cylinders is assumed to be free space.

11.11.10 Energy of a system of charged conductors

11.11.10.1 Energy of a single charged conductor

Consider a single conductor isolated from the electrical effects of all other bodies, the capacitance of the conductor being, say, C .

We have seen that the capacitance of a single conductor can be looked upon as that of a capacitor made of two bodies, of which one is the conductor under consideration

while the other is an infinitely large conductor, as shown in fig. 11-36, with potential zero.

One can work out the energy required in charging the conductor to a given potential, say, V by imagining a process where the charge is built up in infinitesimally small steps. At any stage of this process, let the potential of the conductor be ϕ , when the charge on it is $q' = C\phi$, and let an infinitesimally small amount of charge $\delta q'$ be brought on to the conductor in the process of build-up. Recalling that the potential energy of a charge q placed at a point with potential V in an electrostatic field is qV (see eq. (11-8a)), the energy required for the above small increment of charge of the conductor is $\phi\delta q' = \frac{1}{C}q'\delta q'$. Summing up all these energies, one gets the total energy required to build up the charge from 0 to q , where $q = cV$. This summation reduces to an integration in the limit of $\delta q' \rightarrow 0$, and one obtains the expression for required energy U_E as

$$U_E = \int_0^q \frac{1}{C}q'dq' = \frac{q^2}{2C}, \quad (11-98a)$$

alternative expressions for U_E , the energy of the charged conductor, being

$$U_E = \frac{q^2}{2C} = \frac{1}{2}CV^2 = \frac{1}{2}qV. \quad (11-98b)$$

The suffix 'E' in U_E is indicative of the fact that the energy is electrical in origin.

For instance, in the case of a spherical conductor of radius a , the energy required to charge the conductor with a charge q is (refer to eq. (11-84b))

$$U_E = \frac{1}{2} \frac{q^2}{4\pi\epsilon_0 a}. \quad (11-99)$$

Since the process of charging the conductor results in an electric field being set up around it, the above energy may be interpreted as the *energy associated with the field* itself. This concept of the energy associated with a field is of great relevance, especially

in the context of the *electromagnetic field* (see chapter 14), where the field itself can be considered to be a dynamical system endowed with energy (refer to section 14.4.6.1) and momentum which it can interchange with material particles.

If E be the magnitude of the electric field intensity at any point, say P, in the field, then the expression for the *energy density* of the field at that point turns out to be

$$u_E = \frac{1}{2} \epsilon_0 E^2, \quad (11-100)$$

where it is assumed that the field is set up in vacuum.

This means that, considering a small volume δv around P, the energy associated with the field in this small volume will be $\frac{\epsilon_0}{2} E^2 \delta v$. Thus, the total energy of the field is given by the expression

$$U_E = \int \frac{\epsilon_0}{2} E^2 dv, \quad (11-101)$$

which is obtained by summing up the energies associated with all the small volume elements making up the entire space occupied by the field. This corresponds to a volume integration over entire space in the expression (11-101).

Considering, for instance, the spherical conductor with charge q , the magnitude of the field intensity at a distance r from its centre is given by $E(r) = \frac{q}{4\pi\epsilon_0 r^2}$, and with this expression for E , the integral in eq. (11-101) reduces to

$$\int \frac{\epsilon_0}{2} E^2 dv = \frac{q^2}{8\pi\epsilon_0 a} = \frac{q^2}{2C}, \quad (11-102)$$

where $C = 4\pi\epsilon_0 a$ is the capacitance of the conductor.

Problem 11-15

Check the result (11-102) out.

Answer to Problem 11-15

Here the integration is to be carried out over the volume of space *exterior* to the conductor, since the field intensity is zero in the interior. Imagining a thin spherical shell with radii r and $r + \delta r$ ($r > a$), the volume of the shell is $4\pi r^2 \delta r$ (thickness times area), and the field energy corresponding to the region occupied by this shell is $4\pi r^2 \delta r E(r)$. The volume integral reduces to an integration over r from a to ∞ .

This shows that the expression $u_E = \frac{\epsilon_0}{2} E^2$ can indeed be interpreted as the energy density of the electric field and that the energy required in charging a conductor can be interpreted as the energy associated with the field set up by the charged conductor.

11.11.10.2 Energy of a charged parallel plate capacitor

Consider a parallel plate capacitor as in fig. 11-37(A), (B), but with the condition that the lower plate is earthed at a potential 0 (replacing the earth potential by its actual value V_0 (say) has the effect of modifying the expressions for all other potentials by the addition of V_0 under conditions outlined in sec. 11.12, with any modification of the final result of this section), and that a charge q' has been given to the upper plate at any stage during the process of charging of the capacitor by bringing in infinitesimally small quantities of charge on to it. The potential of the upper plate at that stage is then $\phi = \frac{q'}{C}$ where $C = \frac{\epsilon_0 A}{d}$ is the capacitance (see sec. 11.11.8). The energy required to bring in an additional amount $\delta q'$ of charge is then $\phi \delta q'$. Summing up all these expressions with q' varying from 0 to q , the final charge of the capacitor, one arrives at the expression $\frac{q^2}{2C}$, as in (11-98b), for the energy required to charge the capacitor.

Once again, this energy can be interpreted as the energy associated with the electric field set up by the charged capacitor. For sufficiently large area (A) of the capacitor plates and sufficiently small distance (d) between the plates, the field can be assumed to be confined to the region within the plates and to be a uniform one, its magnitude being given by $E = \frac{q}{\epsilon_0 A}$ (check this out; see sec. 11.11.8). Making use of this expression and multiplying the energy density $\frac{\epsilon_0 E^2}{2}$ with the volume (Ad) occupied by the field, one

obtains the total field energy as

$$U_E = \frac{\epsilon_0}{2} \left(\frac{q}{\epsilon_0 A} \right)^2 A d = \frac{1}{2} \frac{q^2}{\frac{\epsilon_0 A}{d}} = \frac{q^2}{2C}, \quad (11-103)$$

implying once again that the energy required to charge the capacitor can be interpreted as the energy associated with the field set up by the charged capacitor.

In building up the charge of the capacitor from 0 to q by bringing in small increments of charge from infinity to the capacitor plate (the upper one in the present instance), one has to perform work against the earth's electric field as well. If the potential of the earth be V_0 then this additional work, whose amount is $V_0 q$ is stored in the earth's electric field, and does not count as the energy of the field set up between the two capacitor plates owing to the potential difference between these.

11.11.10.3 Electric field energy

More generally, considering a pair of conductors with coefficients of capacitance c_{11} , c_{22} , c_{12} , the energy required to charge these conductors with charges q_1 and q_2 , corresponding to potentials V_1 , V_2 is given by

$$U_E = \frac{1}{2} (q_1 V_1 + q_2 V_2) = \frac{1}{2} c_{11} V_1^2 + c_{12} V_1 V_2 + \frac{1}{2} c_{22} V_2^2, \quad (11-104)$$

and, as in the case of the charged spherical conductor and the parallel plate capacitor, this can be interpreted as the energy of the electrical field set up by the charged conductors, the general expression for the field energy being given by (11-101).

11.11.11 Capacitors in series and parallel

Imagine a pair of parallel plate condensers, one made up of the conducting plates A and B and other of the plates C, D. Suppose the two plates of the first capacitor are given charges q , $-q$, and let the *same* charges (i.e., q , $-q$) be given to the second capacitor as well. The potential difference between the plates of the two capacitors considered

independently of one another will then be $V_1 = \frac{q}{C_1}$ and $V_2 = \frac{q}{C_2}$, where C_1, C_2 stand for their capacitance.

Suppose now that the two capacitors are joined *in series* as shown by the dotted line in fig. 11-40(A), which indicates an electrical connection between the conducting plates B and C belonging to the two capacitors. Since the potentials of B and C before the connection may, in general, differ from one another, there may, in principle, occur a redistribution of charges between the outer and inner surface of the plates. In reality, however, any such redistribution that may occur is negligible, and the charges on the inner surfaces of the plates continue to be as shown in the figure, with only negligibly small amounts of charges appearing on the outer surfaces.

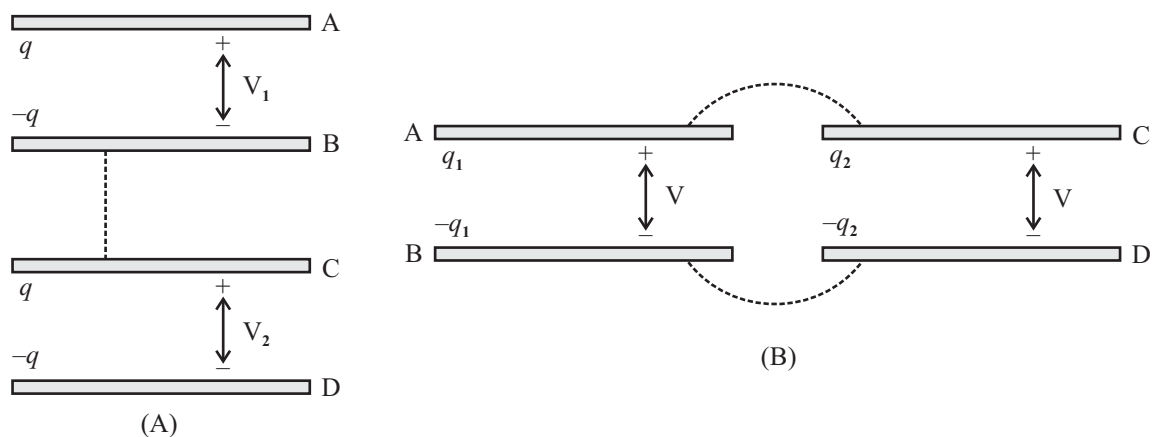


Figure 11-40: Capacitors in (A) series and (B) parallel; in (A), the potential difference between the plates differs for the two capacitors while, under appropriate conditions, the charges producing the fields are the same; in (B), on the other hand, the capacitors have the same potential difference between the plates, but their charges are different.

1. As indicated in sec. 11.11.8.1, small amounts of charges already occur on the outer surfaces of the plates even before the connection.
2. The fact that the charges on the outer surfaces before and after the establishment of the connection are negligibly small (such charges are, however, responsible for the equalization of the potentials between B and C after the connection) can be traced to the smallness of the self capacitances of the plates of either of the two capacitors compared to the magnitude of any of its coefficients of capacitance or,

equivalently, to its capacitance.

The potential difference between A and B and that between C and D therefore remain unchanged, but now B and C are at the same potential, implying that the potential difference between A and D is $V = \frac{q}{C_1} + \frac{q}{C_2}$. Since the charges on A and D are q and $-q$, the two capacitors in series may be looked upon as a single capacitor made up of the plates A and D with a capacitance $C = \frac{q}{V}$. One thus has

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2}, \quad (11-105)$$

giving the *equivalent capacitance* of the two capacitors.

One can arrive at formula (11-105) following another, equivalent, approach. Assume that the two capacitors are uncharged to start with and that the plates B and C are electrically connected as in the figure. If, now, charges q and $-q$ are given to the plates A and D, then in virtue of the usual assumptions (large area of the plates, small distance between them), charges $-q$, q will be induced in the inner surfaces of B and C, while the charges on their outer surfaces will be negligibly small, as will be the charges on the outer surfaces of A and D. This distribution of charges will ensure that the fields inside the materials of the conducting plates will be all zero.

Fig. 11-40(B) shows a pair of parallel plate capacitors with capacitors C_1 and C_2 where now the capacitors are given charges, say, q_1 , $-q_1$, and q_2 , $-q_2$ to start with, such that the potential differences between the plates of the two capacitors considered separately are the same, say V .

Suppose now that the capacitors are connected *in parallel*, with the plates A and C connected to each other and with B and D also similarly connected, as shown by dotted lines in fig. 11-40(B).

Once again, there will be only negligible redistribution of charges between the surfaces

of the conductors (such a redistribution is, in principle, possible because though the potential *differences* are the same, the potentials of A and C need not be the same before the connection and similarly, the potentials of B and D also differ before the connection) and one can look at the capacitors in parallel connection as a single equivalent capacitor with charges $q_1 + q_2$, $-q_1 - q_2$, and with a potential difference V between its plates. Since $C_1 = \frac{q_1}{V}$ and $C_2 = \frac{q_2}{V}$, the equivalent capacitance is given by

$$C = \frac{q_1 + q_2}{V} = C_1 + C_2. \quad (11-106)$$

Equations (11-105) and (11-106) express the rules of *series and parallel connection of capacitors*.

However, these rules, which have been derived with reference to parallel plate capacitors, does *not* apply to capacitors in general (recall our definition of a capacitor as a pair made of *any* two conductors, isolated from the electrical effects of other bodies). For instance, if we have a capacitor made of two arbitrarily chosen conductors A and B, and another with conductors C and D, and if now B and C are connected electrically, then the system of conductors cannot, in general, be looked upon as a single capacitor with capacitance expressed by a formula of the form (11-105), and a similar remark applies to the case of parallel connection as well. This is because of the possible *mutual* electrical effect between the pairs C, D and A, B, which we have tacitly assumed to be negligible in the case of a pair of parallel plate conductors.

The formulae (11-105), (11-106), however, apply to spherical capacitors connected in series or parallel, if the conductors making up one of the capacitors *enclose* those belonging to the other one (the condition of one pair of spherical conductors enclosing the other pair, though sufficient, is, however, not a necessary one, see below). Fig. 11-41 depicts a spherical capacitor made of the conductors A, B enclosed by a second capacitor made up of concentric spherical conductors C, D where, for the sake of simplicity, B, C, D are assumed to be thin spherical shells, each with negligible thickness. Let the radii of the conductors be, respectively, a , b , c , and d , as shown in the figure. Suppose that charges q , $-q$ are given to the conductors A and B making up the first capacitor,

and also to conductors C and D making up the second capacitor before an electrical connection is established between the conductors. If now the conductors B and C are electrically connected then there takes place no redistribution of charges between the conductors. Thus, a charge q is distributed uniformly over the surface of A and also over the outer surface of C, while a charge $-q$ gets distributed uniformly over the inner surface of B as well as that of D.

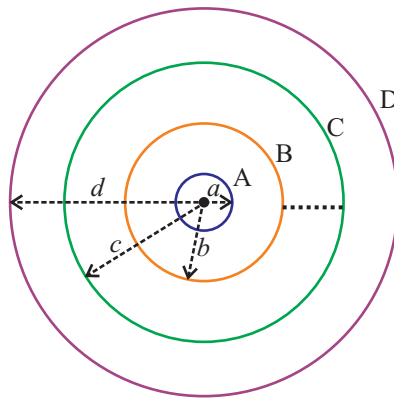


Figure 11-41: Spherical condensers connected in series; A and B make up one of the two condensers with capacitance C_1 , while C and D form a condenser with capacitance C_2 ; on connecting B with C (dotted line) a single equivalent capacitor of capacitance $\frac{C_1 C_2}{C_1 + C_2}$ is formed; the applicability of the series (or parallel, as the case may be) formula to the pair of capacitors depends on the fact that the conductors C, D enclose completely the conductors A, B.

The set-up can be looked upon as a single capacitor made up of A and D, with charges q , $-q$ respectively, and with potentials V (say) and 0 (assuming for the sake of concreteness that the conductor D is earthed), where V can be worked out by the application of Gauss' principle. The equivalent capacitance is thereby found to be given by eq. (11-105), where $C_1 = 4\pi\epsilon_0 \frac{ab}{b-a}$ and $C_2 = 4\pi\epsilon_0 \frac{cd}{d-c}$ are the capacitances of the two capacitors considered separately.

Similar considerations apply in the case of parallel connection of the spherical capacitors too where, it is to be noted, the geometry of the four spherical conductors involved is as shown in fig. 11-41, with the conductors C, D enclosing A, B. It may be mentioned that, while the applicability of the series and parallel formulae to a pair of parallel plate capacitors is only approximate, in the case of a pair of spherical capacitors with the

geometry mentioned above, the formulae are exact.

However, even when the conductors making up one of the spherical capacitors do not enclose the ones making up the other, the series and parallel formulae still apply because of the *shielding* effect of the outer conductor of either capacitor due to which the field external to the capacitor is independent of the charges (q and $-q$) given to the two conductors. What is of crucial relevance here is that, in each of the two spherical capacitors, one of the two conducting spheres encloses the other.

The above considerations relating to series and parallel connections of parallel plate and spherical capacitors, also apply to capacitors made of long coaxial cylinders. Imagine a pair of long coaxial cylindrical conductors forming a capacitance C_1 , and a similar pair of coaxial cylinders forming a second capacitor of capacitance C_2 . In this case the formulae (11-105), (11-106) apply for series and parallel connection of the two capacitors, once again due to the shielding effect of the outer conductor of either capacitor.

More generally, *if the conductors making up any one of the two capacitors have no electrical effect on those making up the other, the series and parallel formulas can be seen to hold.*

11.11.12 Capacitors with dielectrics

In most situations of practical interest, capacitors are formed of conductors with dielectrics filling up the regions between these. Fig. 11-42(A) depicts a parallel plate capacitor with the space between the two plates filled up with a dielectric material with relative permittivity ϵ_r . With charges q and $-q$ on the two plates of the capacitor, the vacuum field is $E_0 = \frac{q}{\epsilon_0 A}$ along the normal to the plates, directed from the positively charged plate to the negatively charged one, and the electric displacement is given by $D = \epsilon_0 E_0 = \frac{q}{A}$. The magnitude of the electric field intensity in the dielectric is then $E = \frac{D}{\epsilon_0 \epsilon_r} = \frac{q}{\epsilon_0 \epsilon_r A}$ (refer to eq. (11-78)), which is simply the vacuum field reduced by a factor of ϵ_r .

The field intensity, which is a uniform one, being the rate of change of the potential

taken with a negative sign, the potential difference between the upper and the lower plates in the figure is given by $V = \frac{qd}{\epsilon_0 \epsilon_r A}$, and the capacitance then works out to

$$C = \frac{\epsilon_0 \epsilon_r A}{d}. \quad (11-107)$$

In other words, the capacitance with the region between the plates filled up with the dielectric is ϵ_r times the capacitance with vacuum ‘filling up’ the region.

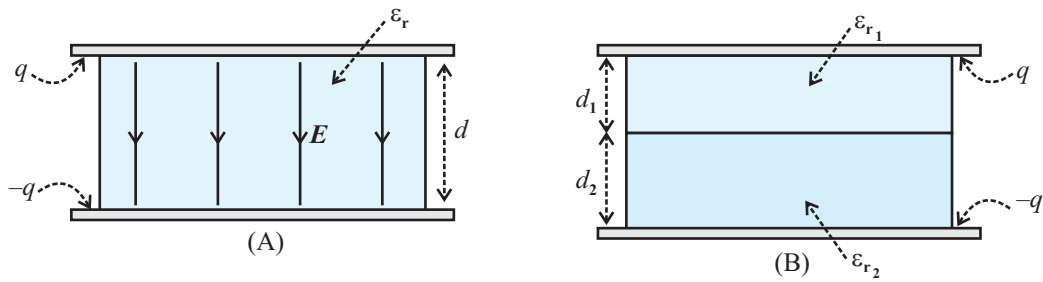


Figure 11-42: Parallel plate capacitors with dielectrics; (A) a single dielectric slab between the plates; (B) two slabs placed end-on-end, making up a series combination.

This rule works for any capacitor formed of a pair of conductors. Thus, considering a spherical condenser with an inner conductor of radius a and an outer conductor of radius b (in the case of the outer conductor being a spherical shell, b stands for its inner radius), with the region between the two conductors filled up with a dielectric of relative permittivity ϵ_r , its capacitance will be given by

$$C = 4\pi\epsilon_0\epsilon_r \frac{ab}{b-a}. \quad (11-108)$$

Fig. 11-42(B) depicts a parallel plate capacitor with the region in between the two plates filled up with *two* layers of dielectrics, one of thickness d_1 ($< d$) of a dielectric of relative permittivity ϵ_{r1} and the other of thickness $d_2 = d - d_1$ of a dielectric of relative permittivity ϵ_{r2} . As shown in the figure, the charges on the two plates of the capacitor are q and $-q$.

One can imagine that the interface between the two dielectrics is formed of a thin conducting plate (without such a plate actually being there) held at the potential of that

plane, where the upper and lower surfaces of the plate are imagined to carry charges $-q$ and q respectively. Such an imagined plate does not change the field in the dielectrics since it does not carry any net charge, but its imagined presence allows us to look at the capacitor as the series combination of two capacitors. Of these, one has a dielectric of thickness d_1 and relative permittivity ϵ_{r1} between its conducting plates, the corresponding parameters for the other being d_2 and ϵ_{r2} .

Making use of the formula (11-107) for the capacitance of a parallel plate capacitor with a single dielectric filling up the region between the two plates and also of the series combination formula (eq. (11-105)) for capacitances, the capacitance C in the present instance is seen to be given by the formula

$$\frac{1}{C} = \frac{1}{\epsilon_0 A} \left(\frac{d_1}{\epsilon_{r1}} + \frac{d_2}{\epsilon_{r2}} \right), \quad (d_1 + d_2 = d). \quad (11-109)$$

The same result can be arrived at without the conducting plate being imagined at the interface between the two dielectrics. For this, one has to work out the field intensities in the dielectrics by making use of the respective electric displacements, and then to evaluate the line integral of the field intensity so as to arrive at the potential difference between the two plates.

Problem 11-16

An air-filled parallel plate capacitor has capacitance 2.5 nF. When the gap between the plates is doubled and, at the same time, a certain dielectric material is inserted so as to fill up the gap, the capacitance increases to 10 nF. What is the relative permittivity of the dielectric?

Answer to Problem 11-16

Since the capacitance is given by the formula $C = \frac{\epsilon_r \epsilon_0 A}{d}$ (see eq. (11-107)), one has $2.5 \times 10^{-9} = \frac{\epsilon_0 A}{d}$ (assuming the relative permittivity of air to be unity), and $10 \times 10^{-9} = \frac{\epsilon_r \epsilon_0 A}{\frac{d}{2}}$, from which the required relative permittivity is seen to be $\epsilon_r = 2$.

11.12 The potential of the earth

When one says that a certain conductor is earthed, we commonly assume that its potential is zero. I have already mentioned that the potential in an electric field is undetermined to the extent of an additive constant. It then seems as if there is no harm in assuming that the earth potential is zero. Indeed, the earth can be looked upon, to a reasonable degree of approximation, as a large conducting sphere whose potential does not change appreciably as other bodies, much smaller than the earth, share charge with it. However, if *this* potential is taken to be zero, then the potential at infinity can no longer be assigned the value zero in working out electrostatic problems. Conversely, taking the potential at infinity to be zero, the earth's potential can no longer be arbitrarily set at zero value.

The electric field intensity, of course, is not subject to such arbitrariness in the choice of a reference potential. The measured value of the field intensity at the earth's surface has been found to be of the order of $100 \text{ V}\cdot\text{m}^{-1}$, directed *radially inward*. Wherefrom does this rather intense electric field come into being?

Any portion of the surface of the earth can be looked upon as one 'plate' of a parallel plate capacitor, the other 'plate' of the capacitor being the ionospheric layer above the earth's surface (looking at the earth as a whole, this corresponds to a large spherical capacitor, with the earth as the inner conducting sphere). The electric field is mostly confined to the space between the ionosphere and the earth's surface, which means that the potential on the other side of the ionospheric layer can be taken to be zero (the same as the potential at infinity). Measuring the rate of variation of the electric field with the altitude, one can then work out the potential of the earth, which turns out to be of the order of several hundred kV. In other words, assuming the potential at infinity to be zero (a common assumption in electrostatics), one arrives at a rather *large* value for the earth's potential.

If, then, one assumes the potential of the earth to be zero in working out any given problem in electrostatics, would it lead to nonsensical results? In reality, for a large

number of problems of practical relevance, it would *not*, introducing only inessential modifications in the results.

For instance, consider the problem of a charged conducting sphere (section 11.11.4) of radius a , with a charge q . If this sphere were far removed from other charged bodies, including the earth and the ionosphere, its potential would be found to be $V = \frac{q}{4\pi\epsilon_0 a}$. This result will, of course, get modified for a charged conducting sphere located near the earth's surface. However, under certain reasonable conditions, the modification will be inessential, at least to a good degree of approximation: the potential simply gets modified to $V = \frac{q}{4\pi\epsilon_0 a} + V_0$, where V_0 stands for the earth's potential. The condition for this result to hold can be expressed as $a \ll h \ll R$, where h denotes the height of the conducting sphere above the earth's surface, and R denotes the radius of the earth.

More generally, if the typical size of the bodies involved in an electrostatic problem and the typical height of these bodies above the earth's surface satisfy the above condition, then all potentials calculated by assuming the earth's potential to be zero get modified in the above manner, i.e., each of the potentials gets increased by V_0 . As a result, the potential *differences* remain almost unchanged.

Thus, considering a situation in which one of the two plates of a parallel plate capacitor (say, the lower one in fig. 11-37) is earthed, its potential will be V_0 and, assuming the self capacitance of either plate to be zero (i.e., considering the limit $\frac{d}{L} \rightarrow 0$), the charge on its outer (i.e., lower) surface will be zero. Let the charge given to the upper conducting plate be q . The self capacitance of the upper plate being assumed to be zero, the charge on the outer surface of this plate will again be zero, and the entire charge q will reside on the inner surface. The charge induced in the inner (i.e., upper) surface of the lower plate will be $-q$, and the potential of the upper plate will be $\frac{q}{C} + V_0$.

As discussed in the present section and in sec. 11.11.8, the condition for these statements to hold is $d \ll L(\sim \sqrt{A}) \ll h \ll R$, where the symbols d , L , h , R have already been defined.

Chapter 12

Electricity I: Steady currents and their magnetic effects

12.1 Electrical Cells

12.1.1 The half cell

A system made up of a conducting *electrode* (commonly, a metal or carbon (graphite) rod) dipped in an electrolyte solution is referred to as a *half cell*. A zinc rod in zinc sulphate solution or a copper rod in copper sulphate constitutes an example of a half cell.

An electrolyte is a compound (commonly, an inorganic salt) that gets split into oppositely charged *ions* in aqueous solution. On the application of an electric field, a current is set up in the electrolyte because of the motion of these ions.

More generally, a material that can carry an electric current by means of ionic transport in the molten state or in a solution, such as sodium chloride, is termed an electrolyte. At times, the term electrolyte is used to refer to the melt or the solution, rather than to the material.

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

Charges of opposite natures appear on the surface of the electrode and in the electrolyte across a thin layer surrounding the electrode in a half cell (fig. 12-1). As the process of charge separation comes to a halt, a state of equilibrium is reached and a potential difference appears between the rod and the electrolyte. The thin region around the electrode across which the potential difference develops is referred to as an electrical *double layer*. The magnitude of the potential difference depends, in general, on the material of the rod, and on the composition and concentration of the electrolyte as also, to some extent, on the temperature of the system.

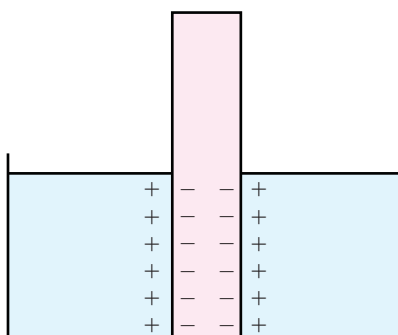


Figure 12-1: A half cell (schematic) made of a conducting electrode dipped in an electrolyte solution; the charge separation between the electrode and the electrolyte across a thin layer surrounding the electrode is shown; a potential difference is developed between the electrode and the electrolyte surrounding it.

If the electrolyte be a solution of the same metal as the one the rod is made of, the system temperature be 298 K, and the concentration of the metallic ion in the solution be 1 mol-dm^{-3} , then the system made up of the rod and the electrolyte is termed a *standard half cell*. The potential difference between the electrode and the electrolyte of a standard half cell is usually measured and expressed in relation to a *standard hydrogen half cell*. For this, an *electrochemical cell* is set up where the standard half cell in question (with a given electrode, say, one made of zinc or copper) is connected with a standard hydrogen half cell through a *salt bridge*, the latter being a device that facilitates the conduction of ions between the two half cells without allowing the electrolytes to get mixed.

1. The standard hydrogen half cell is made up of a platinum electrode dipped in an

acidic solution, commonly an aqueous solution of hydrochloric acid of concentration 1.18 mol-dm^{-3} , kept at 298 K.

2. A salt bridge can be made of a glass tube filled with an electrolyte fixed in a gelatinous medium like agar (derived from varieties of seaweed), or of a filter paper soaked in the electrolyte. A few of the electrolytes commonly used in salt bridges are potassium iodide, sodium sulphate, and potassium chloride. A porous wall separating the electrolytes of two half cells can also serve the same purpose as does a salt bridge.

12.1.2 Electrochemical cell

More generally, an electrochemical cell is formed when two half cells, not necessarily standard ones, are connected through a salt bridge. In some instances, when the same electrolyte is used in the two half cells, a salt bridge is not necessary, and the electrochemical cell is made up of a pair of conducting electrodes dipped in an electrolytic solution.

12.1.2.1 Half-cell potential

The basic processes occurring in an electrochemical cell consists of *oxidation* and *reduction* reactions complementing each other. From a fundamental point of view, a process of oxidation or reduction consists of an *electron transfer*. For instance, a zinc atom (Zn) can give up two electrons to get oxidized into a zinc ion (Zn^{++}), or a hydrogen ion (H^+) can take up an electron to get reduced to a hydrogen atom (H).

When no current passes through an electrochemical cell, and the system made up of the two half cells is in equilibrium, a potential difference exists between the electrode and the electrolyte of each of the half cells, which is characteristic of that half cell and is termed the half-cell potential. The process by which the potential difference is generated in a half cell is that of a charge separation at the electrodes accompanying an oxidation or reduction reaction, as the case may be, that occurs for a transitory period till the equilibrium is established. As the electric field at each electrode resulting from the charge separation builds up, it tends to prevent a further charge separation from

occurring, and a condition of equilibrium finally sets in.

The potential difference of either of the two half cells relative to that of a standard hydrogen half cell can be either negative or positive, depending on whether the tendency of the metal atom to get converted into the ionic form is stronger or weaker than that of a hydrogen atom. Considering two half cells of dissimilar metals, say, A and B, with standard half cell potentials, say, $V^{(A)}$ and $V^{(B)}$ respectively (both being relative to the standard hydrogen half cell), one will have $V^{(A)} < V^{(B)}$ or $V^{(A)} > V^{(B)}$ depending on whether the tendency of atoms of A to form ions is stronger or weaker than that of the atoms of B.

The half cell potentials for zinc and copper, for instance, are respectively -0.76 V and $+0.34$ V, meaning that zinc is more easily oxidized than copper or, put differently, zinc is more *electropositive* than copper.

12.1.2.2 Electromotive force of an electrochemical cell

If, now, an electrochemical cell is formed with the two half cells and equilibrium is allowed to set in (which usually occurs very quickly), then a potential difference will be formed between the two electrodes (which we refer to as electrode A and electrode B respectively), given by

$$V_{AB} = V^{(A)} - V^{(B)}. \quad (12-1)$$

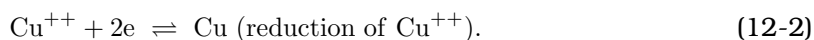
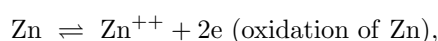
At equilibrium, no current flows through the cell, and V_{AB} , given by the above formula, is then termed its *electromotive force* (*EMF* in short), where A is taken to be the electrode at the higher potential and B the one at the lower potential.

When a current flows through the cell, the potential difference between the electrodes (V_{AB}) is no longer given by $V^{(A)} - V^{(B)}$, i.e., V_{AB} differs from the electromotive force.

12.1.2.3 The Galvanic cell

An electrochemical cell can be used as a *Galvanic cell* (or a *Voltaic cell*, also commonly referred to as an electrical cell) to drive a *current* as required. The electrode at the higher potential is termed the *cathode* and the one at a lower potential the *anode* of the Galvanic cell. When the two terminals of such a cell are connected through a wire (or, more generally, an electrical *circuit*) a current flows through the latter from the cathode to the anode, and energy stored in chemical form in the electrolyte gets converted into other forms, part of which goes to heat up the wire (or the *resistors* (see sections 12.4.2, 12.6.1) in the circuit).

The basic process in a Galvanic cell consists of a set of *oxidation-reduction* reactions, with oxidation occurring at the anode and reduction at the cathode. For instance, in a *Daniell cell*, made up of a zinc half cell (zinc rod dipped in zinc sulphate solution) and a copper half cell (copper rod dipped in copper sulphate solution) connected by a salt bridge, the zinc and copper rods are the anode and cathode respectively. When the two are not connected externally, one has $V_{AB} = 1.1$ V, (where A stands for copper, the cathode, and B for zinc, the anode), the EMF of the cell. Note that this is consistent with (12-1) since, in this instance, $V^{(A)} = 0.34$ V, and $V^{(B)} = -0.76$ V. The basic reactions occurring at the anode and the cathode are, respectively,



When no current flows through the cell, an equilibrium is established in the above two reactions, i.e., their net rate is zero. On the other hand, when the cell delivers a current through an external circuit, the reactions proceed predominantly from the left to the right. Zinc ions pass into the solution, copper atoms are deposited on the cathode, and there occurs a migration of oppositely charged Zn^{2+} and SO_4^{2-} ions through the salt bridge, as a result of which neutrality of the two bulk electrolytes is preserved. It is the EMF of the cell that causes the current to flow in the external circuit, which we will

examine below in greater detail.

Fig. 12-2 depicts schematically a Galvanic cell with half cells of metals A (cathode) and B (anode) connected by a wire, where the voltages (relative to an arbitrarily chosen reference voltage) at various points are shown. The absolute values of the two half cell potentials (more strictly, potential differences) are, respectively, $V_A - V'_A$ and $V_B - V'_B$ while the corresponding potentials relative to the standard hydrogen half cell potential (say, $V^{(H)}$) are

$$V^{(A)} = V_A - V'_A - V^{(H)}, \quad V^{(B)} = V_B - V'_B - V^{(H)}. \quad (12-3)$$

It is important to note that these half cell potentials are *independent* of whether or not a current is flowing through the cell and the external circuit. When the current is zero (i.e., the wire between the cathode and the anode is disconnected) the electrolyte potentials V'_A and V'_B are equal (as we will see below, the difference between the two is proportional to the current), and hence the potential difference between the cathode and the anode is given by

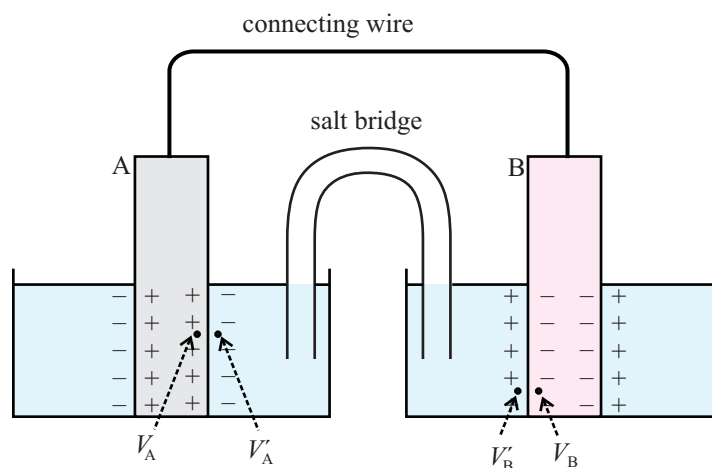


Figure 12-2: A Galvanic cell made of two half cells with electrodes A and B; an external connection between the two electrodes is shown; voltages at various points are indicated; the potential difference between the cathode (A) and anode (B) is $V_{AB} \equiv V_A - V_B$; this equals the EMF $V^{(A)} - V^{(B)}$ of the cell when the current through the cell is zero (i.e., the external connection is cut off).

$$V_{AB} \equiv V_A - V_B = V^{(A)} - V^{(B)}, \quad (12-4)$$

consistent with (12-1). When, on the other hand, a current flows through the circuit, with the current in the electrolytes being directed from the anode to the cathode (in contrast to the current in the external circuit, which flows from the cathode to the anode), one has $V'_B > V'_A$, and then V_{AB} (i.e., the potential difference between the cathode and the anode) drops to a value *less than* the EMF ($V^{(A)} - V^{(B)}$) of the cell.

Two or more Galvanic cells can be connected *in series* to make up a *battery* with an effective EMF larger than that of an individual cell.

Concepts relating to electrical currents and electrical circuits will be in greater details in the following sections.

Problem 12-1

The half cell potential for copper ($\text{Cu}^{2+} + 2e^- \rightleftharpoons \text{Cu}$) with reference to hydrogen is 0.337V, while the EMF of a Cu-Zn electrical cell is 1.100V. If the half cell potential of silver ($\text{Ag}^+ + e^- \rightleftharpoons \text{Ag}$) be 0.800V, what would be the EMF of a Ag-Zn cell? .

Answer to Problem 12-1

HINT: $V^{\text{Ag-Zn}} = V^{\text{Ag-H}} - V^{\text{Zn-H}} = 0.800\text{V} - (V^{\text{Cu-H}} - V^{\text{Cu-Zn}}) = 0.800 - (0.337 - 1.100)\text{V} = 1.563\text{V}$, where the meanings of the symbols are evident.

12.1.2.4 The electrolytic cell

Another variant of the electrochemical cell is the *electrolytic cell* which, in a sense, functions in a manner complementary to the Galvanic cell. While in a Galvanic cell, current is driven in a circuit by making use of chemically stored energy in the electrolyte(s) of the cell, a reverse process takes place in an electrolytic cell, where a current driven

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

through the cell with the help of an external source of EMF (say, another Galvanic cell or battery) causes chemical changes in the form of oxidation-reduction reactions.

Suppose, for instance, that a zinc half cell and a copper half cell are used to make up a Daniell cell, but now the two electrodes of the cell are connected to the two terminals of a battery (fig. 12-3) with an EMF greater than 1.1 V (i.e., the EMF of the Daniell cell when used as a Galvanic cell). The positive terminal of the battery is connected to the copper electrode and the negative terminal to the zinc electrode so that the current, driven by the battery, flows in the external circuit from zinc to copper while, in the electrolytes, the current flows from copper to zinc.

This contrasts with the Galvanic Daniell cell where the current in the external circuit is from copper to zinc and that in the electrolyte(s) is from zinc to copper. Further, in the electrolytic cell of fig. 12-3, the zinc electrode is termed the cathode and the copper electrode the anode, again in contrast to the Galvanic Daniell cell. This way of naming the electrodes ensures that in *both* the Galvanic cell and the electrolytic cell, the current in the electrolyte(s) flows *from the anode to the cathode* (and hence, in the external circuit, from the cathode to the anode)! The basic reactions occurring in the cell are given by (12-2), but now the reactions proceed predominantly from the *right* to the *left*.

Electrolytic cells are in extensive use in the chemical industry. However, these are relevant in the present context only as a class of devices complementary to the Galvanic cells.

On a current is made to flow through an electrolyte, the latter gets split up into ions, and it is the motion of the ions that causes the current to flow through the electrolyte. This phenomenon of dissociation of an electrolyte into ions due to a current flowing through it will be discussed in greater details in section 12.11 below.

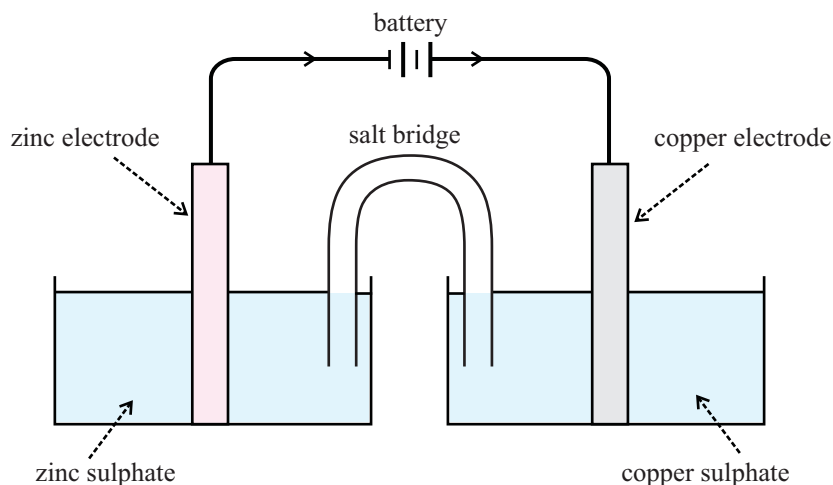


Figure 12-3: A Daniell cell used as an electrolytic cell by connecting the copper and zinc electrodes to the two terminals of a battery, where the copper electrode is connected to the positive terminal and is termed the *anode*, while the zinc electrode is connected to the negative terminal and is termed the *cathode*; the EMF of the battery is to be larger than 1.1 V, the latter being the EMF of a Daniell cell when used as a Galvanic cell; current flows in the external circuit from zinc to copper, and in the electrolytes, from copper to zinc.

12.1.2.5 Primary and secondary cells

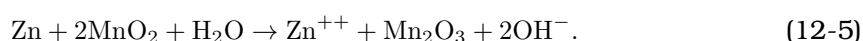
Galvanic cells can be classified into *primary* cells and *secondary* cells (also called *accumulators*, or *storage cells*). A primary cell is one in which the electrolyte gets dissociated during the period the cell supplies current (and energy) to an external circuit. The dissociation products do not remain within the cell in a form appropriate for their recombination into the electrolyte in the original form and hence, after a certain period of delivering energy, the cell becomes useless ('discharged') as a source of EMF.

In the case of a secondary cell, on the other hand, even when the electrolyte gets depleted by the cell supplying current for some period of time, the dissociation products remain available for recombination and the cell can be *recharged* by sending a current from an external source of EMF, during which process it works as an *electrolytic* cell and the electrolyte is regenerated at the cost of energy being supplied to it. Once the recharging process is complete, the cell can once again be used as a source of EMF, i.e., as a Galvanic cell, supplying current and energy to an external circuit.

A familiar example of a primary cell is the *Leclanche dry cell*, where a paste of ammo-

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

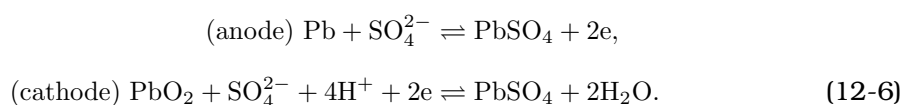
nium chloride and zinc chloride is kept in a cylindrical zinc container in contact with manganese oxide and graphite powder. A carbon rod is inserted along the axis of the cylindrical container and a metal contact is attached to the top of the rod. The carbon rod acts as the cathode and the zinc container as the anode of the cell. The basic reaction taking place in the cell can be written as



Zn^{++} ion passes into the solution at the anode, while electrons pass into the solution at the cathode, being responsible for the production of the OH^- ions. The EMF of the Leclanche cell is 1.5 V.

The 'mercury battery' commonly used in devices such as the electronic watch and the cardiac pacemaker, is made up of mercury dry cells where mercuric oxide (mixed with graphite) is used as the cathode and zinc as the anode. The two are separated by a thin layer of paper or a porous material soaked with an electrolyte, namely, sodium or potassium hydroxide. The EMF (commonly referred to as the 'voltage') of the cell is 1.35 V. As the cell delivers a current, mercuric oxide gets reduced to mercury, and zinc gets oxidized to zinc oxide. However, the use of the mercury battery is a matter of grave environmental concern.

A familiar example of a storage cell is the *lead-acid* cell, several such cells in combination forming a storage *battery*. The lead-acid accumulator is used widely in vehicles, as also in many other areas of practical use. The positive electrode (cathode) of a lead-acid accumulator is made of lead oxide (PbO_2) and the negative electrode (anode), of lead (Pb). The electrodes are immersed in sulphuric acid solution of density 1.25 kg-dm^{-3} . The basic reactions at the anode and the cathode in the cell are, respectively,



When the accumulator works as a Galvanic cell, supplying current and energy to an external circuit, the reactions proceed predominantly from the left to the right, yielding lead sulphate. The EMF of the cell is 1.8 V. As the cell gets discharged with the production of a considerable amount of lead sulphate and its EMF drops, it can be recharged by sending a current through it in the opposite direction from an external source of EMF. During charging, the accumulator functions effectively like an electrolytic cell and the reactions in eq. (12-6) proceed predominantly from the right to the left, thereby regenerating the chemicals in the cell.

A battery made of six lead-acid cells has an EMF of 11 V (approx), and can deliver a current as large as 10 A. However, it cannot keep on delivering this current for an indefinite period since it gets discharged, when the process of recharging has to be initiated.

An important indicator of the performance of such a battery is the duration for which the cell can go on supplying a given magnitude of current, and is commonly expressed in *ampere-hours*. For instance, if the battery is capable of supplying a current of 2 A for 50 hours before getting discharged, then it can be said to have a capacity of 100 ampere-hours.

A few rechargeable batteries in common use other than the lead-acid battery are the nickel-cadmium battery (the use of which is in question because of the toxicity of cadmium), the nickel-metal hydride battery, and the lithium-ion battery.

Rechargeable dry batteries are in wide use in industrial and home appliances.

12.2 Electrical conductors and electric current

Materials around us can be classified as *conductors*, *semiconductors*, and *insulators* in terms of their electrical properties. Of these, the conductors and the semiconductors are capable of carrying electrical *currents*, while a current cannot ordinarily be set up in an insulator. We will, for now, be concerned with conductors alone.

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

The atoms and molecules in a conductor are, in general, arranged in a regular pattern making up a *crystalline structure*. As a consequence of such regular arrangement, the electrons of the conductor do not remain bound to individual atoms, but can move about throughout the volume of the crystalline material, behaving like *delocalized* electrons. Among these, one group of electrons act as electrical *carriers* whose states can be changed by the application of a weak electric field, leading to the setting up of a current. The remaining electrons are kind of 'frozen' into their respective states that are not affected significantly by the application of a weak electric field, and these cannot act as carriers of an electric current.

In a *semiconducting* material, the number density of carriers happens to be less than that in a conductor. Moreover, a special feature of a semiconductor as compared to a conductor is that, apart from electrons, there occur positive *holes* as carriers in it though, in the ultimate analysis, holes are explained in terms of states of an assembly of electrons. On the other hand, the number density of carriers in an *insulator* is negligible. You will find all this discussed in greater details in sections 11.10 and 19.2.6.

When an electric field of even a low strength is applied in a conductor, the carriers, which are negatively charged electrons, gain energy from the field, acquiring an acceleration in a direction opposite to that of the field. The momentum of the electrons cannot, however, go on increasing indefinitely as a result of this acceleration because they get *scattered* from the vibrating atoms in the crystalline material, thereby losing their acceleration. What happens as a result of the two competing processes of acceleration and scattering is that eventually a *steady state* is arrived at when the electrons acquire a uniform *drift velocity* in a direction opposite to that of the applied electric field.

If one looks at any one electron in particular, one will find that, due to the collisions with the vibrating atoms in the crystal, that electron moves about in a manner resembling a *random* motion, depending on the temperature. However, when observed over a period of time, it will be found to have a *net* motion, i.e., motion on the average, in the direction mentioned above (fig. 12-4). This average motion of the electrons in a direction opposite

to that of the applied field, is referred to as a drift motion.

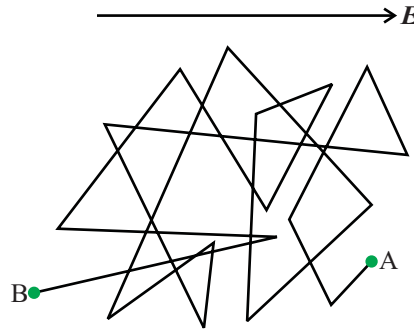


Figure 12-4: Illustrating the drift motion of a negatively charged carrier (electron) in an applied electric field (E); as a result of successive collisions with the atoms of the crystalline material, the carrier appears to follow a random course, but after a large number of collisions, it suffers a net motion in a direction opposite to that of the applied field, moving from the point A to the point B.

1. When I speak of looking at any one electron in particular, I make a simplification that is not really valid. It is not possible, even in principle, to distinguish one carrier from another among the multitude of carriers. The identity of individual carriers is lost in the collection of all the carriers in a conductor. Thus, strictly speaking, the drift velocity is to be defined by looking at the average motion of a large number of carriers taken together.
2. The atoms and molecules making up the crystalline structure do not remain fixed at their respective positions. At any given temperature, they vibrate about their respective mean positions in a random manner. This motion is referred to as 'thermal vibration', since the mean energy of the vibration increases with the temperature of the material. In spite of this thermal vibration, however, the mean positions of the atoms form a regular crystalline arrangement. As the temperature of the conductor is increased, the thermal vibration becomes more vigorous, resulting in more frequent scattering of the carriers from these vibrating atoms, and a consequent decrease in the drift velocity of the carriers.

Thus, when an electric field is set up in a conductor, there results a drift motion of the carriers, and hence a flow of charge. Since the carriers in a conductor are the negatively charged electrons, there occurs a flow of negative charges in a direction opposite to that

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

of the electric field. One describes this by saying that a *current* is set up *in the direction* of the electric field. In other words, the direction of current is by convention defined as that in which *positive* charges would flow under the influence of the applied electric field.

In a semiconductor, on the other hand, the current is set up not only by negatively charged carriers but by positively charged ones as well. The direction of the current is still the one in which positive charges tend to move under the action of the applied electric field. This implies that the currents due to the negative and positive carriers add up.

The quantitative measure of the flow of charge in the neighborhood of any given point in a conductor is expressed in terms of the *current density* at that point. This is illustrated in fig. 12-5 where a small planar area is shown around a point P, its plane being perpendicular to the direction of the electric field intensity E , and hence to the direction along which the drift motion of the carriers takes place. Suppose that the area of the planar element is s , and the magnitude of the drift velocity of the carriers in a very small neighborhood of the point P is v .

The figure shows an imagined cylindrical volume erected on the planar area as base, where the axis of the cylinder is parallel to the direction along which the drift motion occurs, its length being v . Let the number of carriers per unit volume in the conductor near P be n , and let e denote the magnitude of the charge of a carrier. The total charge of the carriers included in the cylinder is then $nevs$, and this amount of charge will cross the area s per unit time in the direction of the drift motion.

In other words, the *rate of flow of charge per unit area* around P is nev . Because of the negative charge carried by the electrons, this flow takes place in a direction opposite to that of the field intensity (recall that here e and v have been defined to be positive quantities, which is why a separate mention of the direction of drift motion is necessary).

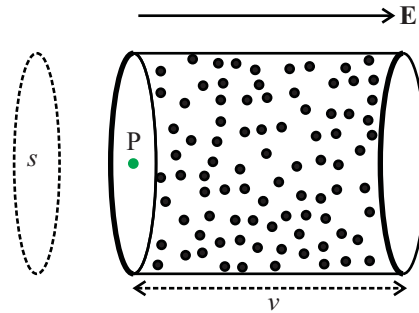


Figure 12-5: Carriers (schematic) in an imagined cylinder based on an area element around a point P in a conductor; v is the drift velocity of the carriers in a small neighborhood of the point P; all the carriers in the cylinder cross the area s in unit time due to their drift motion caused by the electric field E ; the resulting flow of charge per unit area per unit time gives the current density at P.

Assuming that the applied electric field is sufficiently weak, i.e., its magnitude (E) is small, the magnitude of the drift velocity v may be assumed to be proportional to E , the magnitude of the field strength:

$$v = \mu E, \quad (12-7)$$

where the constant μ is termed the *mobility* of the carriers. The current density then works out to

$$j = ne\mu E, \quad (12-8)$$

where the direction of the current density is, by definition, along the applied field. When the electric field intensity and the current density are looked upon as three dimensional vectors, the expression for the current density is seen to be

$$\mathbf{j} = ne\mu \mathbf{E}. \quad (12-9a)$$

An alternative expression for the current density, in terms of the drift velocity vector, is

$$\mathbf{j} = ne\mathbf{v}, \quad (12-9b)$$

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

where e is now the *actual* charge of a carrier (i.e., possesses negative for an electron) and \mathbf{v} is the drift velocity vector whose direction is opposite to that of \mathbf{E} in the case of negatively charged carriers. It is convenient to make use of this expression for the current density in order to define the conductivity (refer to formula (12-9c) below). However, the mobility is defined to be a positive quantity and hence, the symbol e in expressions (12-9a) and (12-9d) stands for the magnitude of the charge of a carrier.

The quantity $ne\mu$ is commonly denoted by σ , and is termed the *electrical conductivity* (or, in brief, the conductivity) of the conductor at the point P. The relation between the current density and the electric field intensity can then be expressed as

$$\mathbf{j} = \sigma \mathbf{E}, \quad (12-9c)$$

where

$$\sigma = ne\mu. \quad (12-9d)$$

This is a basic formula for describing the current in a conductor. The reciprocal of the conductivity is referred to as the *resistivity* (ρ):

$$\rho = \frac{1}{\sigma} = \frac{1}{ne\mu}. \quad (12-10)$$

1. In a *homogeneous* conductor the conductivity σ is the same at all points, and its value depends on the material of the conductor as also on the temperature. Moreover, the conductivity may depend on other physical factors as well, like the presence of a magnetic field and the presence of electromagnetic radiation.
2. With an increase in the temperature of the conductor, the rate at which the carriers get scattered by the vibrating atoms of the crystalline structure also increases. As a result, the drift velocity for any given strength of the applied electric field decreases, with a corresponding decrease in the mobility of the carriers. In other words, the conductivity of a conductor decreases with an increase in the temperature. For a semiconductor, on the other hand, while the mobility decreases with

an increase in the temperature, there occurs an increase in the number density of the carriers, which causes the conductivity to *increase*, overriding the effect of the decrease in mobility.

3. What I have presented above is an outline of how a *macroscopic* description of the drift motion of the carriers relates to a *microscopic* description involving motions of individual carriers in the electric field. The two are reconciled only when an average is taken over the motions of a large number of carriers. In the following, it will be the macroscopic description that I will stick to, with only occasional references to the microscopic picture. In this, when I refer to the motion of a single carrier, I will actually mean a carrier in the *average* sense, i.e., one that is imagined to follow the average motion of a large number of carriers. In *this* description, there is no room for the collisions and scatterings that a carrier undergoes as an individual particle. What remains is just a uniform drift motion under the applied electric field. The kinetic energy of the carrier remains constant as long as it moves under a uniform field.
4. The above paragraphs include a few basic concepts of an essentially *classical* theory of the electrical conductivity of materials. The classical theory, however, has a number of drawbacks in explaining the salient features relating to electrical conductivity. For instance, it cannot explain in quantitative terms the observed temperature dependence of the conductivity over a substantial temperature range. Still, the concepts outlined above provide useful ingredients to a correct, *quantum* theory of the conductivity.
5. The formula (12-9c) applies to an *isotropic* conducting material, where the current density has, by definition, the same direction as \mathbf{E} . Conductors are made up of crystalline materials and a single crystal is, in general, a *non-isotropic* medium. However, a conductor commonly consists of a large number of micro-crystals packed with random orientations, as a result of which it behaves effectively as an isotropic medium. In the case of a crystalline medium, the current density (which is defined by assuming that the carriers are positively charged) is no longer in the direction of the applied field, and the conductivity is, in general, a *tensor*, that can be represented in terms of a 3×3 symmetric matrix.

Problem 12-2

A potential difference of $V = 2\text{V}$ set up across a conducting wire of length $l = 1.0\text{m}$, produces a current density $j = 1.5 \times 10^6 \text{A}\cdot\text{m}^{-2}$ in it. Work out the resistivity of the material of the conductor. If the number density of carriers be $n = 6.0 \times 10^{27} \text{m}^{-3}$ what is the mobility of the carriers?

Answer to Problem 12-2

HINT: The electric field strength in the conductor is $E = \frac{V}{l} = 2.0\text{V}\cdot\text{m}^{-1}$. The resistivity is obtained from the relation $\rho = \frac{1}{\sigma} = \frac{E}{j}$, i.e., $\rho = 1.33 \times 10^{-6}$ SI units ($\Omega\cdot\text{m}$, see sec. 12.4 below). Then, making use of formula (12-10), one gets $\mu = \frac{1}{ne\rho} = \frac{1}{6.0 \times 10^{27} \times 1.6 \times 10^{-19} \times 1.33 \times 10^{-6}}$ SI units, i.e., $7.8 \times 10^{-4} \text{m}^2\cdot\text{s}^{-1}\cdot\text{V}^{-1}$.

12.3 The current set up by a Galvanic cell

Fig. 12-6 shows a Galvanic cell whose terminals (electrical contacts on the two electrodes) have been connected with a conducting wire. Before connecting the wire to the terminals of the cell, there was no current flowing through it, and every small volume element in it was electrically neutral, the positive and negative charges in any such element canceling each other. Recall that the positive charges in the wire are not mobile like the electrons - they are much heavier, and are capable only of thermal vibrations about their respective mean positions. In contrast, the electrons are mobile and acquire a drift motion under the influence of an electric field.

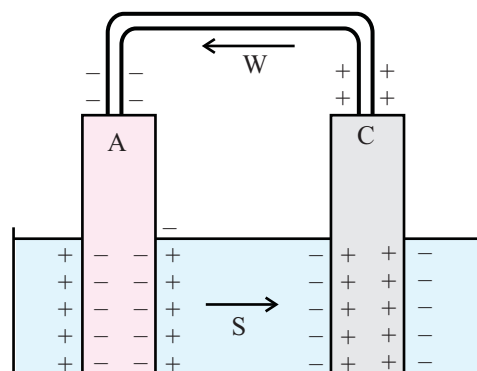


Figure 12-6: Steady current in a wire connecting the terminals of a Galvanic cell: unbalanced static charges and the flow of the steady current have been indicated; A, C, S, and W represent respectively the anode, the cathode, the electrolyte solution, and the conducting wire.

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

As I have already mentioned, there takes place a charge separation between each of the electrodes and the electrolyte in its immediate vicinity. The electrodes thereby get charged with opposite charges, and an electric field is set up in the region surrounding the electrodes. The electric field is very strong in the electrical double layers around the two electrodes, and is much weaker in other regions of space, but is nevertheless there.

When the wire is connected between the terminals of the cell, the mobile electrons in the wire acquire a drift motion under the influence of this electric field, whereby the electrons in the wire close to the anode terminal drift towards the positively charged cathode. The charge neutrality of the volume elements of the wire is thereby violated, as a result of which local electric fields are created and electrons move into the wire from the anode while a similar flow occurs from the wire into the cathode. At the same time, the equilibrium of the electrical double layers in the cell tends to be upset. Electrochemical reactions in the cell are thereby initiated so as to bring back the equilibrium in no time, and the process gets repeated - electrons flow from the anode into the wire and from the wire into the cathode, while the electrical double layers remain almost in equilibrium, moving away from it by small degrees and then tending to move back quickly. To a very good degree of approximation, the electrical double layers can be assumed to remain in their equilibrium configurations while, at the same time, oxidation and reduction processes continue in the regions near the two electrodes.

Eventually, but almost in no time, a *steady state* is established in the entire system. In this steady state, chemical reactions occur in the cell (oxidation at the anode and reduction at the cathode) at a constant rate, and at the same time, a drift motion of carriers occurs in the wire, resulting in a steady current flowing through it *from the cathode towards the anode*. All the while the electrical double layers remain in their equilibrium states, i.e., the states that one observes in the respective half cells.

Immediately after the connection between the two electrodes is established by means of the wire, there occurs a transitory non-steady state preceding the establishment of the steady state. In this non-steady state, an electric field is set up in the wire and the flow of carriers, which initially is confined to the regions close to the electrical contacts

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

is extended into regions away from the contacts, eventually leading to a steady drift motion of carriers throughout the wire and to the setting up of a steady electric field in it. This transitory process of setting up of a steady electric field in the wire proceeds by means of an *electromagnetic wave* traveling down the wire. The electromagnetic wave travels with a very high velocity - close to the *velocity of light in vacuum* (see chapter 14 where you will find ideas relating to electromagnetic waves discussed in greater details). This is why the steady state is brought about within a vanishingly short time.

In the steady state, all physical quantities relating to the system made up of the wire and the cell remain constant in time. For instance, for a homogeneous wire the charge density remains constant at the value zero everywhere in its interior while, at the same time, a flow of carriers occurs along it at a constant rate. Imagining a small volume element in the wire, the rate at which carriers enter into the volume element through its boundary surface must then be the same as the rate at which they leave it (fig. 12-7).

Recall that in an *electrostatic field* set up by a system of charges in equilibrium, the charge density is similarly constant in time everywhere in space but, *there does not occur any flow of charges*. In other words, a situation involving a steady current, while apparently resembling that relating to an electrostatic field, is, in reality, quite distinct from the latter. What needs a special mention in this context is that, the electric field intensity in the interior of a conductor is everywhere zero in the electrostatic condition (refer back to section 11.10.2), while the field intensity in the wire carrying the steady current *has to be* non-zero since there is a non-zero current density everywhere in the wire (refer to eq. (12-9c)). This electric field in the wire is directed *from the cathode towards the anode*.

In the steady state, *a non-zero charge density appears on the boundary surface* of the wire and, under certain conditions, in its interior, and remains constant in time. It is this distribution of charges along the wire that is responsible for the electric field in the wire, the latter causing the current to flow in it. Recall that, in the electrostatic condition, the charge density in the interior of a conductor has to be zero, and all the

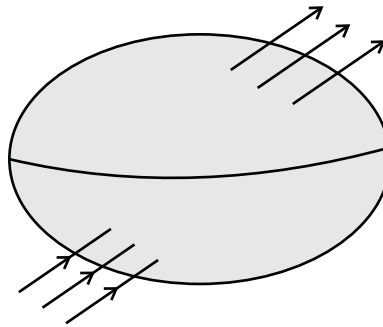


Figure 12-7: Illustrating the drift motion of carriers into and out of an imagined volume element of a conducting wire carrying a steady current; as much charge moves into the volume element in any given time as leaves it through the boundary surface of the volume element; the charge density at every point remains constant in time.

unbalanced charge of a conductor resides on its boundary surface.

By contrast, there can, in principle, be a non-zero charge density even in the interior of a wire carrying a steady current, provided the physical characteristics of the wire (such as its conductivity) vary from point to point. For a *homogeneous* wire, on the other hand, the charge density in the interior turns out to be zero, thus resembling the electrostatic condition.

The charge surface charge distribution on a conductor in the static situation is such as to lead to a vanishing electric field intensity in the interior of the conductor. By contrast, the surface charge distribution in a homogeneous wire carrying a current (for which the volume charge density happens to be zero) is to be such as to lead to a uniform electric field in it.

In this context, I have to mention the setting up of a *magnetic field* in the region surrounding the conducting wire and the electrical cell (you will find a primer on magnetic fields set up by steady currents later in the present chapter). Indeed, a current flowing in a conductor causes the appearance of a magnetic field around the conductor. If the current happens to be steady, then the strength of the magnetic field also remains constant in time.

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

However, in the transitory non-steady condition prior to the steady state being arrived at, both the electric and magnetic field strengths keep on changing with time, and an electromagnetic wave propagates along the boundary of the wire.

According to the electromagnetic theory, if the strength of the magnetic field remains constant in time, then the electric field turns out to be a *conservative* one. This means that one can define an *electrical potential* at every point in the field. If V_P and V_Q be the potentials at any two points P and Q in the field, then the work required to move a particle of charge q from Q to P is given by the expression $q(V_P - V_Q)$, which does *not* depend on the *path* along which the charge is transferred from Q to P (refer back to section 11.4.2). Knowing the value of the potential at all points in the field, one can work out the field strength everywhere. From this point of view, the electric field associated with a steady current resembles an electrostatic field, both being conservative in nature.

However, the potential associated with a steady current in a wire set up by means of an electrical cell is, in a sense, only a notional one, since the potential varies so rapidly in the electrical double layers in the cell as to be, to all intents and purposes, a *discontinuous* one. This is a consequence of the fact that the cell acts as a source of EMF.

Employing the concept of the potential for the electric field associated with the steady current in the wire connecting the terminals of a Galvanic cell, I will present a few interesting conclusions relating to the transformation of energy in section 12.5.1.

12.4 Ohm's law. Electrical units

12.4.1 Current density and current

With the concept of current density introduced earlier (sec. 12.2), we define the electrical *current* through a small surface of surface area δs around a given point P (say) as

$$\delta I = j\delta s, \quad (12-11)$$

where the surface is assumed to be perpendicular to the direction of the current density vector \mathbf{j} at P. It signifies the amount of charge crossing the area δs in unit time. If the plane of the small surface is not perpendicular to the direction of \mathbf{j} , and the normal to the surface makes an angle θ with \mathbf{j} , then the expression for the current through it, i.e., the amount of charge crossing the surface per unit time, will be $j\delta s \cos \theta$ (reason this out), or, in other words,

$$\delta I = \mathbf{j} \cdot \delta \mathbf{s}, \quad (12-12)$$

(see fig. 12-8) where, among the two possible unit vectors normal to the small area element under consideration, $\delta \mathbf{s}$ is defined in terms of the one that points in the general direction in which the current actually crosses the element (see fig. 12-8; recall that the direction of flow of current is defined to be opposite to the direction of drift velocity of negative carriers). The current I through an arbitrarily specified surface, not necessarily of small area, is defined by imagining it to be made up of a large number of small parts and summing up the currents through all these small parts. Once again, the current through any given surface signifies the charge crossing the surface per unit time.

The unit of current in the SI system of units is 'ampere' (A, in short), which is defined operationally in terms of the force between two currents (see section 12.8.1). This gives the unit of current density as $\text{A}\cdot\text{m}^{-2}$, since current density is simply current per unit area. Again, since current is charge (crossing a surface) per unit time, the unit of charge is the product of the unit of current and that of time, i.e., A·s, which is given the

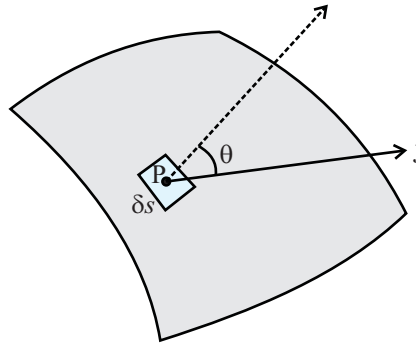


Figure 12-8: Defining the current through a given surface; the surface is imagined to be divided into a large number of small area elements, where each element can be assumed to be a planar area; one such element is shown, around a point P on the surface, where the current density is \mathbf{j} , and the normal to the area element (dotted line) makes an angle θ with the direction of \mathbf{j} ; the current through the area element in the direction of the normal is then given by the expression $j \cos \theta \delta s$ or, in other words, by eq. (12-12); the current I through the surface under consideration is then obtained by summing up the currents through all the area elements making up the surface; it signifies the charge crossing the surface per unit time.

name *coulomb* (C in short). While we have introduced this unit earlier in the context of electrostatics, the coulomb is defined operationally in the SI system with reference to the ampere.

Recall that we have also introduced the volt as the unit of electrical potential (or, more specifically, potential difference) in chapter 11 as $\text{J}\cdot\text{C}^{-1}$, which thus relates the volt with the coulomb, and thus with the ampere, through the unit of work or energy, joule (J, in short). Note that the volt is not a basic unit in the SI system, but is a derived one, related to the basic units A, m, kg and s.

The unit of electric field intensity can be defined in terms of the volt through equation (11-16) where we find that this unit is just the volt multiplied with the unit of length with exponent -1 , i.e., $\text{V}\cdot\text{m}^{-1}$. The unit of electrical conductivity is then obtained (refer to eq. (12-9c)) as $\text{A}\cdot\text{m}^{-1}\cdot\text{V}^{-1}$.

12.4.2 Resistance and resistivity

I will now define the *resistance* of a conductor and introduce the unit of resistance as ‘ohm’ (Ω). For this, refer to fig. 12-9 which shows a homogeneous conductor (or a portion

thereof) of a cylindrical shape with uniform cross-section A and length l , carrying a steady current I , the potential difference between the two ends (A and B) being V . Let us assume that the values of all relevant physical quantities in the situation under consideration are constant in time.

Since the current enters into the region of the conductor shown in the figure through the cross-section A and leaves it through B, and no current crosses the remaining boundary surface (S), the current density j and the electric field intensity E are uniform along the length of the conductor, both being directed from A to B. One then has

$$I = jA, \quad E = \frac{V}{l}, \quad (12-13)$$

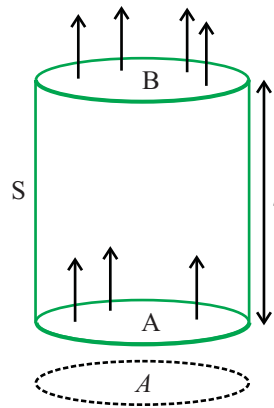


Figure 12-9: Introducing the idea of *resistance*; a conductor (or a part thereof) is shown, in the form of a uniform cylinder of length l and area of cross section A ; current enters and leaves the conductor through the end faces A and B; the current density and electric field intensity are uniform everywhere; the current I through the conductor is proportional to the potential difference V between the end faces, the constant of proportionality - the resistance of the conductor - being, in the given instance, given by eq. (12-16).

(refer to equations (12-11) and (11-19) where we assume, moreover, that the current density is constant throughout any cross section of the conductor. Then, substituting in equation (12-9c), we obtain the following relation between the current through the

conductor and the potential difference between the ends A and B:

$$V = \frac{l}{\sigma A} I, \quad (12-14)$$

where σ is the electrical conductivity of the material of the conductor, given by the formula (12-9d) (recall that $\rho = \frac{1}{\sigma}$ is the resistivity of the material, see formula (12-10)).

The quantity $\frac{l}{\sigma A}$ is termed the *resistance* of the conductor. It depends on the physical characteristics of the conductor as also on its dimensions, being proportional to the length l and inversely proportional to the area of cross section A .

Denoting the resistance of the conductor (or a portion thereof, the one under consideration) as R , we rewrite the relation (12-14) as

$$V = IR. \quad (12-15)$$

I have derived this relation between the current through a conductor and the potential difference between its ends in the case of a homogeneous conductor of uniform cross section, under the assumption that the current density does not vary across the cross-section. The conclusions to be drawn from this formula are twofold: (i) the current (I) through the conductor is *proportional* to the potential difference (V) where the proportionality constant, the *resistance* of the conductor, depends on the resistivity of the material, as also on the dimensions - for a *given* conductor under given physical conditions, the current changes in proportion to the potential difference, a phenomenon that goes by the name of *Ohm's law*; and (ii) the dependence of the resistance on the conductivity or resistivity of the material and on the dimensions of the conductor in the particular situation depicted in fig. 12-9 is of the form

$$\begin{aligned} R &= \frac{l}{\sigma A} \\ &= \rho \frac{l}{A}, \end{aligned} \quad (12-16)$$

which we obtain by combining eq. (12-14) with eq. (12-10).

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

How much of above remains true when the conductor is *not* of uniform cross section, or not homogeneous, and if the current density varies in direction and magnitude across the cross-section of the conductor? To start with, the first of the two conclusions above remains true, i.e., the proportionality between I and V - Ohm's law - continues to hold, and the constant of proportionality is termed the resistance of the conductor regardless of whether it is homogeneous or whether it is of uniform cross section.

However, one then will have to be more careful in defining I and V . Imagine two thin wires to be attached to two points of the conductor and a circuit to be completed by connecting the wires with the terminals of an electrical cell, there being no other electrical cell or source of EMF connected to the conductor. In this situation, if V be the potential difference between those two points, and if the current I enters and leaves the conductor through the two thin connecting wires, then I and V will be found to be proportional to each other.

If one thinks of the potential difference as the 'cause', and the current as the 'effect' (a potential difference between two ends of a conductor *causes* a current to flow through it) then one can paraphrase Ohm's law by saying that the *effect is proportional to the cause*.

On the other hand, the second of the above two conclusions falls through: the resistance can no longer be expressed in the simple form (12-16). For a homogeneous conductor the resistance can still be expressed as the resistivity ρ times a factor that depends on the dimensions, but this factor can no longer be expressed in the form $\frac{l}{A}$. Indeed, for a conductor of arbitrary shape, finding an expression for the resistance generalizing equation (12-16) is a tall order. A further complication comes in if the conductor is not homogeneous because then the resistivity varies from point to point, and one single resistivity cannot be taken out as a factor in the expression for R , as in eq. (12-16).

It is easy from eq. (12-16) to work out the unit for resistance in terms of that for conductivity. But things get a bit involved here if we keep on trying to express everything in terms of the basic units m, kg, s and A. Instead, it is simpler to introduce a derived

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

unit, called the 'ohm' (Ω), for resistance keeping in mind though, that it can ultimately be related to the above four basic units (work it out). One can then express the units of conductivity and resistivity in terms of the ohm and the other three basic units kg, m, and s, the results being $\Omega^{-1} \cdot m^{-1}$ and $\Omega \cdot m$ respectively.

Notice from formula (12-16) that, of two wires of given material having the same length, the thinner wire will have a greater resistance compared to the thicker one. This is why windings of transformers handling large currents use thick wires since these windings need to have a small resistance so as to prevent the winding from getting heated excessively (see section 12.5.4 for the formula for the rate of heating of a current-carrying conductor). For the same reason, fuse wires in lines required to carry large currents are required to be thick since a thin wire gets heated even for a relatively small current and melts.

Looking at formula (12-16) one also observes that, of two wires made of the same material and having the same cross-section, the shorter wire will have a smaller resistance compared to the longer one. This is why connecting wires in high precision electrical measuring circuits are required to be short and thick since otherwise their resistance would introduce errors in the results of measurement.

An electrical *circuit* is a set-up including a number of electrical devices where the latter may include sources of EMF (such as an electrical cell) and conducting bodies, characterized by their respective resistances, through which there can occur a flow of current. The conducting bodies, such as wires of various length and thickness, are termed *resistors*. At times, a circuit may include an air gap created in an otherwise closed current path, where the flow of current gets broken by the air gap. The circuit (or part thereof) is then said to be *opened* by the introduction of the air gap. One may look at the air gap in an open circuit as a resistor of *infinitely large* resistance. In practice, however, even such an air gap possesses a large but finite resistance, as a result of which the current through it can be assumed to be negligibly small.

Problem 12-3

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

A conducting wire of cross-section $A = 10^{-6} \text{ m}^2$ is of length $l = 0.2 \text{ m}$, and has a resistance $R = 1.0 \Omega$. If the number density of free electrons in the material of the wire be $n = 2.0 \times 10^{28} \text{ m}^{-3}$, and if a potential difference $V = 5.0 \text{ V}$ is maintained between its ends, calculate (a) the current density, assuming it to be uniform across the cross-section of the wire, and (b) the mobility of the electrons.

Answer to Problem 12-3

HINT: In accordance with Ohm's law, the current in the wire is $I = \frac{V}{R} = 5.0 \text{ A}$. Hence the magnitude of the current density is $j = \frac{I}{A} = 5.0 \times 10^6 \text{ A}\cdot\text{m}^{-2}$. This can be expressed as $j = nev$, where $e = 1.6 \times 10^{-19} \text{ C}$ is the magnitude of the charge of an electron, and v is the drift velocity of the electrons. Thus, $v = \frac{j}{ne} = 1.56 \times 10^{-3} \text{ m}\cdot\text{s}^{-1}$ (approx). The magnitude of electric field intensity in the wire causing the flow of current is $E = \frac{V}{l} = 25 \text{ V}\cdot\text{m}^{-1}$. The mobility is then obtained as $\mu = \frac{v}{E} = 6.24 \times 10^{-5} \text{ m}^2\cdot\text{s}^{-1}\cdot\text{V}^{-1}$.

Problem 12-4

A resistor is in the shape of a cylinder with a slight taper, where the end faces of the cylinder are of cross sections $A_1 = 1.5 \times 10^{-6} \text{ m}^2$ and $A_2 = 1.3 \times 10^{-6} \text{ m}^2$, and where its length is $l = 0.50 \text{ m}$. If the mobility of the free electrons in the material of the cylinder be $\mu = 1.0 \times 10^{-4} \text{ m}^2\cdot\text{s}^{-1}\cdot\text{V}^{-1}$, and their number density be $1.5 \times 10^{28} \text{ m}^{-3}$, estimate the resistance of the cylinder, stating any reasonable assumptions that you need.

Answer to Problem 12-4

HINT: We follow the basic ideas in sections 12.2 and 12.4, and make the assumption that the component of the current density parallel to the axis of the cylinder at any given point in it is constant over the cross section of the cylinder through that point. Denoting by j the value of this component of the current density for a cross section of the cylinder with area A at a distance, say x from the end with cross section A_1 (which we assume to be the end at a higher potential, say V compared to the other end, assumed to be held at potential 0), the current crossing this cross section will be $I = jA$.

Since, in the steady state, there cannot be any accumulation of charge anywhere, I has to be the same at all such cross sections, which means that j has to increase from the wider to the narrower

end of the cylinder. Since $j = ne\mu E$ ($e = 1.6 \times 10^{-19}$ C), the electric field intensity along the axis of the cylinder has to increase as well. Considering a small slice of the cylinder of thickness δx , one can write $E = -\frac{\delta V}{\delta x}$ or, in other words, $I = jA = -ne\mu A \frac{\delta V}{\delta x}$. Imagining the cylinder to be made up of a large number of such slices, and taking into consideration that I has to be constant for all such slices, one can write $I = -ne\mu \sum \frac{\delta V}{\delta x}$, where the summation is over all the slices making up the cylinder. In other words, $I = ne\mu \sum \frac{V}{\delta x}$, where the summation reduces to an integration ($\int \frac{dx}{A(x)}$) in the limit of the thickness of each slice tending to zero (the integral $\int dV$ is $-V$).

This gives $V = IR$, where the resistance R is given by the expression $R = (ne\mu)^{-1} \int \frac{dx}{A(x)}$, and where the area $A(x)$ at a distance x from the wider end can be expressed as $A(x) = A_1 + \frac{A_2 - A_1}{l}x$. Working out the integral and making use of given values, $R = (ne\mu)^{-1} \frac{l}{A_1 - A_2} \ln \frac{A_1}{A_2} \approx (ne\mu)^{-1} \frac{l}{\frac{A_1 + A_2}{2}} = 1.49 \Omega$ (approx).

12.4.3 Temperature dependence of resistivity

The electrical conductivity (and hence the resistivity) of a material depends on its temperature, principally through the temperature dependence of the mobility μ of the carriers of electrical current (see formula (12-9d)). Consequently, the resistance of a resistor made of the material also depends on the temperature. If R_0 be the resistance of the resistor at some given reference temperature T_0 (commonly taken to be 273 K or, for practical purposes, around 300 K), and R its resistance at any other temperature T , then one can express R as a function of T by a linear relation of the form

$$R = R_0(1 + \alpha(T - T_0)), \quad (12-17)$$

where α is, to a good degree of approximation, a constant, independent on the choice of T_0 and T , and is referred to as the *temperature coefficient* of resistance of the resistor under consideration (and hence also of the resistivity of the material of the resistor; the change in the dimensions of the resistor due to the change in temperature can be ignored to a good degree of approximation). The temperature coefficient of resistivity of conductors are all positive, which means that the resistance of a conductor increases with temperature. This contrasts with the temperature coefficients of resistivity of *semi-*

conductors, which are found to be negative.

The theory underlying the relation of the resistivity of a material (including its temperature dependence) to its basic material properties is more involved than that indicated in sec. 12.2, where only the *classical theory* of conductivity was briefly introduced. On looking at the details of the theory and its correspondence with observed facts, a number of fundamental deficiencies are detected, which are remedied only when *quantum* considerations are included. For instance, the explanation of a rather pronounced variation of the temperature coefficient of the resistivity of a conductor at very low temperatures, and of the emergence of *superconductivity* is essentially dependent on quantum theoretic considerations.

If the resistance R of a resistor gets changed by δR when the temperature (T) is changed by ΔT , then formula (12-17) can be written as

$$\delta R = \alpha R \Delta T. \quad (12-18)$$

It may be mentioned that, in numerous situations of practical interest, α can not only be taken to be a constant, independent of the temperature T but, moreover, its value turns out to be *small* in the sense that, even for a value of δT of considerable magnitude one can write, to a good degree of approximation, $\alpha \Delta T \ll 1$, which implies $\delta R \ll R$.

12.5 Steady current in a conductor produced by an electrical cell

12.5.1 Transformation of energy

The conducting wire and electrolyte in the electrical cell together make up a *closed circuit* through which there occurs a flow of current. Let us consider points P_1, P_2, \dots, P_6 on a closed path in this circuit as shown in fig. 12-10. Of these, P_1 and P_2 are located at the two ends of the wire, close to the cathode and the anode terminals respectively, while P_3, P_4 , and P_5, P_6 are across the double layers at the anode and the cathode. During the

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

flow of the steady current through the wire, suppose that a carrier of charge $-q$ ($q > 0$; recall that the carriers in a conductor are negatively charged) follows the closed path from P_2 to P_1 , and then through P_6, P_5, P_4, P_3 , back to P_2 .

In reality, it is unlikely that one single carrier ever moves through such a closed path.

What is of greater relevance is the average motion of a large number of carriers. However, considering a closed path followed by a single carrier allows for a conceptual simplification which I want to make use of.

We will assume that during the flow of the steady current, the electrical double layers around the two electrodes remain in equilibrium, and that the potentials at P_1 and P_6 are the same, as are the potentials at P_2 and P_3 . The potentials at P_1 (P_6), P_5 , P_2 (P_3), and P_4 are then nothing but the ones denoted by V_A, V'_A, V_B, V'_B respectively in fig. 12-2. As mentioned earlier, one has $V_A > V_B$ and the quantity $((V_A - V'_A) - (V_B - V'_B))$ represents the electromotive force of the cell.

In this case, the motion of the carrier from P_2 to P_1 occurs from a higher to a lower potential energy ($-qV_B$ to $-qV_A$), but this decrease in the potential energy *does not* appear as an increase in the kinetic energy of the carrier. Instead, the kinetic energy of the average drift motion remains constant. What eventually happens is that the decrease in the potential energy appears as an increase in the *internal energy* of the material of the wire, which gets heated up. The vibrations of the atoms and molecules about their mean positions become more vigorous, these vibrations being of a random nature at the level of individual atoms. The energy associated with the vibrations is sometimes referred to as the *thermal energy* of the material of the wire.

At the microscopic level, the energy of a carrier tends to increase in vanishingly small installments under the influence of the electric field, but every time the carrier gains a little in kinetic energy, it suffers a collision with some vibrating atom or other in the crystal structure, and gives up the kinetic energy to that atom, whereby the average kinetic energy of the carrier remains unchanged during the drift motion. As a result,

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

there occurs an increase in the energy of thermal vibration of the atoms and molecules making up the crystal structure.

Since the potentials at P_1 and P_6 are the same, there does not take place any energy exchange between the carrier and the surrounding material as the former travels from P_1 to P_6 . In reality, however, each of the electrodes gets heated to a small extent, but that does not have much significance in the present context.

We now consider the motion of the carrier from P_6 to P_5 . In this part of the journey, the motion of the carrier cannot be described as a drift motion, but is of a different kind, where it moves from a *higher* to a *lower* potential. Here the carrier is transferred from one molecule (or group of atoms) to another in a *chemical reduction type reaction*. Indeed, since $V'_A < V_A$ (recall that the difference $V_A - V'_A$ is the half cell potential at the cathode; commonly, however, the half cell potential is measured with reference to the hydrogen half cell potential), the electrical force on the carrier acts in a direction *opposite* to its direction of motion, i.e., in other words, the potential energy of the carrier *increases* from P_6 to P_5 by the amount $q(V_A - V'_A)$. On the other hand, the mean kinetic energy remains unchanged, and the question arises as to where the increase in the potential energy comes from ?

The source of this energy is the energy of *chemical bonding* in molecules or groups of atoms, or even the binding energy of an electron to an ion. The above amount of energy gets released in the reduction reaction at the cathode and 'pushes' the carrier past the electrical double layer from P_6 to P_5 .

The next phase of the journey of the carrier from P_5 to P_4 occurs from a lower to a higher potential. During this phase, there once again occurs a decrease in the potential energy of the carrier, now by an amount $q(V'_B - V'_A)$, where $V'_B > V'_A$ - this last being the condition for the flow of a current through the wire and the electrolyte. This energy is, in the ultimate analysis, transferred to the electrolyte, which gets heated up. Once again, the kinetic energy of the carrier remains unchanged since the velocity of drift motion

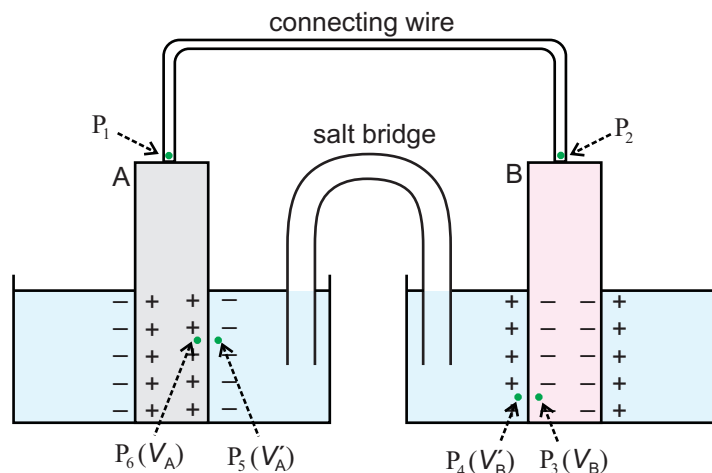


Figure 12-10: Points P_1, \dots, P_6 chosen on an imagined closed path in the circuit made up of the electrolyte in an electrical cell and a connecting wire; A and B denote the cathode and the anode respectively (see fig. 12-2) and the potentials at the various points are indicated; a carrier, imagined to follow the closed path, experiences changes in its potential energy while its kinetic energy of drift motion remains unchanged; at the same time, the carrier receives energy released in the chemical reactions occurring at the two electrodes of the cell; in a complete circuit made by the carrier, the gain and loss in potential energy balance each other, and the net result is the conversion of the energy made available by the chemical reactions to the thermal energy of the wire and the electrolyte, whereby these two get heated up.

remains constant for a steady current.

Next, in moving from P_4 to P_3 , the carrier travels from a lower to a higher potential energy as in the case of its journey past the potential barrier at the cathode surface. The increase in potential energy equals $q(V'_B - V_B)$, where $V_B - V'_B$ is the half cell potential at the anode. The supply of this energy comes once again from the energy of chemical bonding, which gets released in the oxidation reaction taking place at the anode and pushes the carrier past the electrical double layer there.

Finally, the carrier moves from P_3 to P_2 without any exchange of energy with its surroundings, since the potentials at P_3 and P_2 are equal (in reality, however, the two potentials differ slightly, and the anode gets heated to a small degree).

As the carrier makes a complete circuit, the following energy changes occur: (a) The potential energy of the carrier attains back its initial value (check that the changes in potential energy occurring in the various parts of the circuit indeed add up to zero); (b)

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

the internal energy of the wire and the electrolyte(s) increases by the amount $(q(V_A - V_B) + q(V'_B - V'_A))$, causing these to get heated up; and (c) the oxidation and reduction processes in the electrical double layers release energy amounting to $(q(V_A - V'_A) + q(V'_B - V_B))$. The equality of the latter two amounts, where both are equal to q times $V^{(A)} - V^{(B)}$ (i.e., q times the EMF of the cell), is in accordance with the principle of conservation of energy. Evidently, what happens here is simply a *transformation* of energy.

1. While I have spoken of collisions of a carrier with the atoms or molecules making up a crystalline structure, I do not really mean collisions with single atoms, one at a time. A better way of putting things would be to say that the carrier gets scattered from *all* the atoms of the crystal taken together, since the atoms are coupled or *bound* to one another by forces resembling those exerted by springs. Thus, a carrier does not interact and exchange energy with an atom in isolation, but does so with all the atoms taking part in the interaction collectively.
2. What is relevant in a collision suffered by a carrier in a crystalline environment is the fact that the arrangement of the atoms at any given instant of time departs from a perfectly regular one because of their thermal motion. A perfectly regular arrangement would not impede the motion of the carrier, and it is the deviation from perfect regularity that causes the scatterings suffered by it. Another major cause of irregularity in a crystalline structure relates to the presence of *faults* that inevitably arise during the formation of the crystal and its subsequent history where it experiences various *stresses* and *damages*. The carrier gets scattered from these faults, once again with the net result of heating up the material.
3. The collisions experienced by a carrier in a *liquid* environment can, in a sense, be looked upon as ones where there is a *preponderance of faults*, since there occurs a short range correlation between the atoms (or groups of atoms) in a liquid, but such correlations are broken on a longer scale.

The energy of chemical bonding released in one complete circuit of a carrier has been seen to be $q\mathcal{E}$ where $\mathcal{E}(= V^{(A)} - V^{(B)})$ is the EMF of the cell. However, the whole of this energy may not appear in the form of thermal energy of the connecting wire and the electrolyte, since some part of it may be transformed to other forms and may be used for various different purposes. The energy made available by the cell as it supplies

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

a current through a circuit is sometimes referred to as *electrical energy*, and one can convert this energy into various other forms, as necessary.

For instance, if the wire be a coil *suspended in a magnetic field*, and if the coil be capable of turning about an axis, then the flow of current through it may result in a rotational motion of the coil, causing it to act as a *motor*. In the latter case, part of the energy released in the chemical reactions at the two electrodes, is used up in maintaining the rotational motion of the coil. This principle has an enormous range of applications in present day technology.

12.5.2 The pathway of energy flow

We have seen in sec. 12.5.1 that the principle of energy conservation is obeyed in the process of the flow of a steady current in the circuit made up of a resistor and an electrical cell. Thus, while there occurs an increase in the internal energy of the wire and the electrolyte, this increase is made up for by a decrease in the chemical energy stored in the electrolyte. In other words, there occurs a transformation of energy during the flow of current. However, the above analysis does not tell us anything of the *pathway* of energy flow. How, in other words, does the energy released in the chemical reactions occurring at the two electrodes get handed down to the material of the wire and to the electrolyte in the cell?

A detailed answer to this question needs a lengthy analysis involving concepts in the theory of electromagnetic fields. I will therefore give here only a brief description of the pathway of energy flow, without telling you how exactly the pathway makes possible the transformation of energy in the circuit.

1. The potential energy of the carriers is nothing more than an accounting concept that lets us keep track of the work done on these by the forces operative in the system. In the present context, these forces are all of electromagnetic origin. In other words, the heating up of the wire and the electrolytes(s) must, in the ultimate analysis, result from the transfer of energy by means of the electromagnetic field. This is the pathway that we are now concerned with.

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

2. To my knowledge, this question of transformation of energy and the pathway of energy flow in a circuit carrying a current has not yet been answered in all details. A good introduction to the problem, with an outline of a solution, will be found in [11], a paper by I. Galili and E. Goihbarg ('Energy transfer in electrical circuits: a qualitative account'), in American Journal of Physics, vol. 73, p 141-144 (2005). In the present section, I have, in the main, followed these authors. Another introduction to the same problem is to be found in [5] (R.P. Feynmann, R.B. Leighton, and M. Sands, 'Feynman Lectures in Physics', Narosa Publishing House Reprint, New Delhi, 1995, vol 2, sec. 27-5).

As I have mentioned earlier, the flow of current is associated with the development of charges on the electrodes and on the conducting wire, and an associated electric field in and around the wire and the electrolyte. At the same time, a *magnetic* field is also set up in the region of space containing the wire and the electrolyte because of the flow of current (see section 12.8.3).

A result of great relevance in *electromagnetic theory* (see chapter 14) states that the simultaneous presence of electric and magnetic fields in any given region of space leads to the *transport of energy* through that region, the rate of flow of energy per unit area at any given point being termed the *Poynting vector* at that point (see section 14.4.6.1).

In the set-up involving an electrical cell and a resistor connected between the terminals of the cell, it is the *flow of energy by means of the electromagnetic field* (generated due to the simultaneous development of electric and magnetic fields) that provides for the pathway through which energy is transferred from the electrical double layers to the material of the wire and to the bulk of the electrolyte.

In fig. 12-11, the directions of the Poynting vector at various points in the region of space around the electrical cell and the resistor are shown with thick arrows. One obtains the directions of the Poynting vector by looking at the electric and magnetic field intensities at these various points, and then making use of the expression (14-10) for the Poynting vector. While drawing the figure, I have made a number of simplifications, where A and B represent the electrodes of the cell C, the two dotted lines demarcate the electrical

double layers near the electrodes, and the resistor is denoted by the rectangular block R. The resistor is assumed to be connected to the terminals of the cell by means of connecting wires of *zero resistance*, which means that these wires are not heated up, and no energy transfer takes place into these wires.

From the figure, one observes that electromagnetic energy flows *out* from the regions around the two electrodes in the cell, and the direction of flow of energy changes from point to point, with the result that the energy flows *into* the resistor and the bulk of the electrolyte, causing the heating up of these bodies. While looking at the figure you have to keep in mind that the regions of the cell around the electrical double layers are distinguished from the bulk of the electrolyte(s) by the direction of the electric field (recall that it is not the electric field that drives the current in the double layers, but the release of chemical energy).

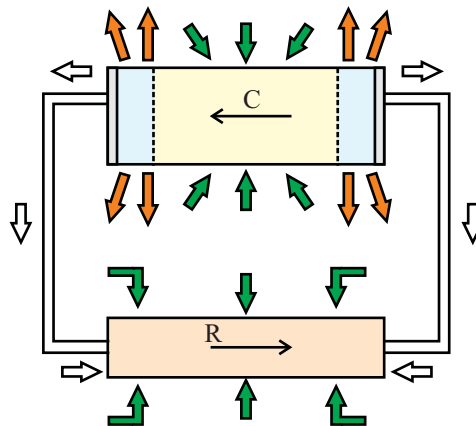


Figure 12-11: Pathway of energy flow in a circuit made up of a cell (C) and a resistor (R), along with connecting wires; the directions of the Poynting vectors at various points in and around the circuit are shown, from which the energy pathway can be deduced; the bulk of the electrolyte in the cell is distinguished from the electrical double layers at the two ends of the cell (not to scale); energy flows out from the two ends and into the resistor as also into the bulk of the cell; for the sake of simplicity, the connecting wires are assumed to have negligible resistance, as a result of which there is no flow of energy into these wires.

12.5.3 Electromotive force (EMF) and source of EMF

The energy released in chemical reactions in an electrical cell causes the cell to send a current in an external circuit (as also through the electrolyte of the cell itself). The cell acts here as a *source of EMF*. The EMF of the source is defined to be the energy delivered by it as a *unit* quantity of charge is made to traverse a closed path from the cathode to the anode and then back to the cathode through the electrolyte at the cost of this energy.

In section 12.5.1 we imagined a carrier with charge $-q$ to traverse a closed path through the points P_2, \dots, P_6 and back to P_2 shown in fig. 12-10). If, instead of the charge $-q$, a positive charge of unit magnitude is considered, then one can similarly imagine it to follow a closed path, but now in the reversed order. The energy released in chemical reactions in the process can be obtained in a similar manner as above, and is given by the expression $(V^{(A)} - V^{(B)})$, which has already been referred to above as the electromotive force of the cell and has been denoted by the symbol \mathcal{E} . It has been seen to equal the potential difference between the cathode and the anode (V_{AB}) in a situation when the cell does not deliver any current, and its unit is volt (i.e., $\text{J}\cdot\text{C}^{-1}$). It is determined by the half cell potentials at the cathode and the anode, being equal to their difference, and hence depends entirely on the internal structure of the cell.

Notice that the EMF can also be expressed as $(V_A - V_B) + (V'_B - V'_A)$, i.e., as the sum of the potential difference between the cathode-end and the anode-end of the wire and that between the anode-end and the cathode-end of the electrolyte. According to Ohm's law, both these quantities depend on the current (I) flowing through the circuit (though their *sum* does not, since it is completely determined by the half cell potentials at the two electrodes). If, for instance, the wire connecting the two terminals of the cell were replaced with one having different resistance, then the two potential drops would be different.

Supposing that the resistance of the connecting wire is R and that of the electrolyte is

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

r , one has, according to Ohm's law, the relations

$$V_A - V_B = IR, \quad (12-19a)$$

$$V'_B - V'_A = Ir. \quad (12-19b)$$

Summing up the above two expressions, one obtains

$$\mathcal{E} = I(R + r), \quad (12-19c)$$

where \mathcal{E} stands for the EMF of the cell. The current delivered by the cell is thus related to its EMF as

$$I = \frac{\mathcal{E}}{R + r}. \quad (12-19d)$$

If the external connection between the electrodes is broken then one can assume the resistance R to be infinitely large (corresponding to the resistance of an insulator - the insulator does not allow any current to flow through it regardless of what the potential difference across it is) and hence one has $I = 0$ (eq. (12-19d)) which implies (eq. (12-19b)) $V'_A = V'_B$, and $V_{AB} = \mathcal{E}$, as I have already mentioned.

For a current I set up in the circuit by the cell, the potential difference V_{AB} between the cathode-end and the anode-end of the connecting wire (denoted by V below for the sake of brevity) is given by the two equivalent expressions

$$V = V_A - V_B = IR = \frac{R}{R + r}\mathcal{E}, \quad (12-20a)$$

and

$$V = \mathcal{E} - Ir. \quad (12-20b)$$

As seen from eq. (12-20b), the potential difference between the ends of the external

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

resistance connecting the two terminals of the cell is *less* than the potential difference between the cathode and the anode in the *open circuit* condition (i.e., when $I = 0$, this being simply the EMF of the cell) by an amount Ir , the potential drop between the anode-end and the cathode-end of the electrolyte in the cell. This potential drop internal to the cell is sometimes referred to as ‘lost volts’.

Is the electrical cell the *only* means of setting up of a current in a circuit? In reality, a current can be set up by means of devices other than a Galvanic cell, these being *sources of EMF* of a different nature compared to the electrical cell. Of especial importance among these are the ones making use of the electromotive force resulting from *electromagnetic induction* (see section 13.2). Fig. 12-12 depicts a closed loop of wire, where a magnetic field is set up in a direction perpendicular to the plane of the wire (the magnetic lines of force (see sec. 12.8.7) are shown with double-headed arrows). If the strength of the magnetic field is made to increase or decrease with time then it is found that a current is set up in the wire. Notice that the wire is *not* connected here with the terminals of an electrical cell. This means that the source of energy making possible the flow of current in the wire is of a different kind.

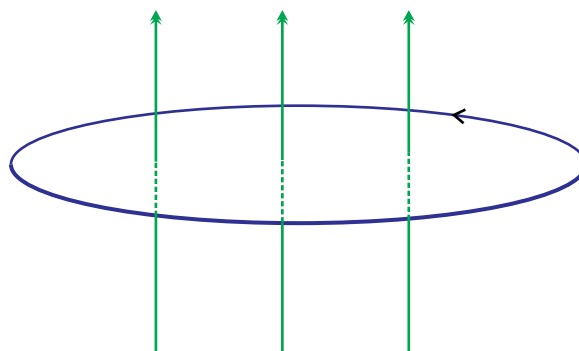


Figure 12-12: Magnetic field (with field lines depicted with double-headed arrows) set up in a direction perpendicular to the plane of a closed loop of wire (concepts relating to magnetic fields are introduced later in the chapter); if the strength of the magnetic field is made to change with time, an electromotive force is set up in the wire loop, causing a current to flow through it; this is the phenomenon of *electromagnetic induction*; the arrow on the wire loop shows the direction of the current that would result if the strength of magnetic field is made to *decrease* with time.

Assume, for instance, that the magnetic field has been created with the help of a per-

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

manent magnet (magnetic fields can also be created with the help of electromagnets which act as temporary magnets). Then, the increase or decrease in the magnetic field strength can be effected by drawing the magnet towards or away from the wire loop, which requires a mechanical force to be exerted on the magnet, and hence some work to be done on the system from outside. It is this work which serves as the source of energy setting up an EMF in the closed circuit made up of the wire loop. The work required to make a unit charge traverse the loop of wire along a closed path is, by definition, the EMF (\mathcal{E}) generated by the changing magnetic field. The work required to make q amount of charge traverse the closed path along the loop is then given by

$$W = q\mathcal{E}. \quad (12-21)$$

This formula holds, in general, for any closed circuit with a source of EMF active in it. Since the current (I) is the rate of flow of charge, the rate of supply of energy by the source of EMF is given by

$$\mathcal{P} = I\mathcal{E}. \quad (12-22)$$

Here \mathcal{P} is referred to as the *power* supplied by the source of EMF in sending the current I in the circuit.

This definition of emf in a circuit can be generalized to one according to which an EMF may be associated with a closed path, not necessarily a path lying in one or more conducting materials. Imagining the closed path to be made up of a large number of small segments, and denoting the vector length of a typical segment by $\delta\mathbf{r}$, if \mathbf{E} denotes the electric field intensity at any point within this segment, the work done on a charge q to make it move from one end of the segment to the other will be $q\mathbf{E} \cdot \delta\mathbf{r}$. The total work required to make the charge make a complete traversal of the closed path can then be expressed in the form of an integral $q \oint \mathbf{E} \cdot d\mathbf{r}$ taken over the closed path. Comparing with eq. (12-21), the emf associated with the closed path is given by the integral $\oint \mathbf{E} \cdot d\mathbf{r}$.

According to this definition, the EMF associated with a closed path is non-zero only if

the electric field intensity, considered as a vector field, deviates from being a conservative one because, the line integral of a conservative vector field along a closed path is necessarily zero. Put differently, a non-zero value of the EMF implies that the electric field cannot be derived from a potential. Even when a potential is used to describe the electric field as its directional derivative (or *gradient*) with a negative sign, that potential cannot everywhere be a uniquely defined, or single-valued, function of position. For instance, in the case of closed circuit including an electrolytic cell, the potential approaches different limiting values as a double layer is approached from either side.

In this book, you will find an occasional non-uniformity of notation relating to the electromotive force. While the EMF has been denoted by the symbol \mathcal{E} above, it has elsewhere been denoted by the symbol E to make the equations look simpler and also for the sake of conformity with common usage. The meaning of the symbol E as an EMF as distinct from its other possible meaning as the magnitude of an electric field intensity will, in most instances, be clear from the context.

12.5.4 Heating effect of current: Joule's law of heating

As seen above, the energy imparted to the conducting wire in a complete circuit made by a charge $-q$ is $q(V_A - V_B)$, which we write in brief as qV , where V denotes the potential difference between the cathode-end and the anode-end of the conducting wire. The same amount of energy may be seen to be imparted if a charge $+q$ makes a complete circuit, but now in the opposite direction, i.e., in the direction of the current. This implies that the *rate* of supply of energy to the material of the conducting wire for a current I flowing through it is given by

$$P = IV, \quad (12-23a)$$

where V stands for the potential difference between the two ends of the conducting wire.

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

Using Ohm's law (eq. (12-15)), this can be written in the alternative forms

$$P = \frac{V^2}{R} = I^2 R. \quad (12-23b)$$

Similar considerations tell us that the rate of energy released to the electrolyte is given by

$$P' = vI = \frac{v^2}{r} = I^2 r, \quad (12-24)$$

where $v (= V_B' - V_A')$ stands for the potential difference between the anode-end and the cathode-end of the electrolyte. Clearly, one has

$$\mathcal{P} = P + P', \quad (12-25)$$

which is an expression of the fact that the rate at which energy is supplied by the cell is the sum of the rates at which energy is imparted to the conducting wire and to the electrolyte. More generally, if a part of the energy supplied from a source of EMF is converted to some other form (say, to work performed on a coil in imparting kinetic energy to it or to maintain its rotational motion) at the rate P'' then formula (12-25) will have to be modified to

$$\mathcal{P} = P + P' + P''. \quad (12-26)$$

An alternative form of the above formula is

$$\mathcal{P} = P_1 + P_2, \quad (12-27)$$

where $P_1 (= P + P')$ is the rate at which heat is produced in the entire set-up including the conducting wire and the electrolyte (or, more generally, in all the resistances making up the circuit), sometimes referred to as the rate of energy *dissipated* in it, and P_2 (the same as P'' in eq. (12-26)) is the rate at which energy is expended on other heads.

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

1. In the case of light being emitted from a coil in the circuit as in an incandescent lamp, the energy emitted as light is to be included in P_1 since it comes from the heating of the lamp coil.
2. While talking of the rate of energy dissipation in a material due to a current flowing through it, I have referred to a situation where the current is a steady one. More generally, for a time dependent current $i(t)$, the instantaneous rate of energy dissipation is given by the expression $P(t) = i(t)^2 r$, where r is the resistance of the conductor under consideration.

Any of the formulas in eq. (12-23a) and eq. (12-23b) expressing the rate of heating of a resistance R when a current I flows through it (with the potential difference between the two ends of the resistance being V) is referred to as *Joule's law of heating*.

Problem 12-5

A heating coil has a potential difference $V = 5$ V applied across it, where the resistivity of the material of the wire is $\rho = 2.0 \times 10^{-8} \Omega \cdot \text{m}$, and its cross-section is $A = 2.0 \times 10^{-6} \text{ m}^2$. If the rate of energy dissipation in the coil be $W = 2000$ W, what is its length (l)?

Answer to Problem 12-5

HINT: The rate of energy dissipation is given by the formula $W = I^2 R = \frac{V^2}{R}$, where I stands for the current through the coil and R for its resistance. The latter is given by the formula $R = \rho \frac{l}{A}$. Combining the two, one gets $l = \frac{AV^2}{\rho W}$. Using given values, $l = 1.25$ m.

12.5.5 Summary: electrical cells, EMFs, and currents

Let me summarize what we have had so far relating to the flow of steady current in a resistor whose ends are connected to the two terminals of an electrical cell.

An electrical double layer is set up at the surface of each electrode in contact with the electrolyte where ions or electrons are added to or removed from the electrolyte, and a potential difference is created across the double layer. The charge separation and the

potential difference across the double layer is a consequence of chemical changes at the interface. This causes a potential difference to arise between the electrodes, due to which there occurs a drift motion of free electrons in the conducting wire, tending to neutralize this potential difference. The latter, however, is renewed continuously due to ongoing chemical changes at the double layers. Though the electrolyte gets gradually depleted due to the chemical changes, we can ignore this depletion in the time scale over which the current remains substantially steady.

While the drift motion of the electrons continues in the conducting wire, the net charge density in its interior remains zero (assuming for the sake of simplicity that the material of the wire is homogeneous), though the electric field intensity in the wire remains at a non-zero value. This intensity is accounted for by a stationary charge density at the boundary surface of the conducting wire (there may, however, arise a stationary non-zero volume charge density if the conductor is inhomogeneous and the conductivity is not uniform). One can speak in terms of a unique value of the potential at every point inside the conductor, meaning thereby that the field inside the conductor is a conservative one as it should be in the case of a field generated by static charges. On the other hand, the field is *not* of a conservative nature when one takes into consideration the electrical double layers since the potential approaches two distinct values as one approaches a double layer from two sides.

In a more complete description, one can include the chemical processes at the electrodes along with the setting up of the electric current in the conductor and work with the *electrochemical potential* of the electrons in the conductor as also of the ionic species in the electrolyte. The chemical reactions and the current then constitute a *steady non-equilibrium* process in the entire system made up of the conductor, the electrodes, and the electrolyte. However, we will not take up such a description in this book. While we assume that a steady current is set up in the conducting wire, we will not enter into the detailed considerations regarding the underlying processes.

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

Though there occurs a steady flow of current in the conductor, at a microscopic level there occurs a drift motion of the electrons *as also* a random motion. During this motion the electrons in the conductor collide with vibrating atoms and structural imperfections in the conductor. What happens in the process is that the electrons tend to gain kinetic energy in moving through the electric field, but continually give up this energy to the material of the conducting wire in small steps through collisions. The conductor is thereby heated up, and the energy imparted to the material of the conductor is made up for from the chemical bonding energy released at the electrical double layers.

A similar process occurs in the electrolyte in between the two electrical double layers where the carriers of current are the ionic species rather than the electrons, and the heating occurs in the electrolyte itself. At a microscopic level, as the ions move through the electrolyte under the influence of the electric field set up in between the two electrodes, their kinetic energy tends to increase, but the latter is given away to the electrolyte itself by means of collisions. This loss of energy to the electrolyte is once again made up for by the release of chemical bonding energy. As energy is converted from one form to another, the entire process conforms to the principle of conservation of energy.

The electrical cell is not the only means for the setting up of a current in a conductor. Other devices can also act as *sources of EMF*, supplying the energy necessary for the setting up of the current. The EMF (\mathcal{E}) operating in a closed circuit is defined as the energy supplied by the source as a unit charge is made to move through the closed circuit by it. In general, a source is characterized by an *internal resistance* (r), in terms of which the current set up in a conductor of resistance (R) is given by formula (12-19d), while the potential drop across the resistance is given by (12-20b) or, equivalently, by the last expression in eq. (12-20a). The rate of supply of energy by the source is given by the formula (12-22). Either the whole or a part of this is lost as heat in the various resistances making up the circuit while part of it can be converted to other forms of energy like, say, the kinetic energy of rotation of a coil about an axis.

12.6 Series and parallel combination of resistances

12.6.1 The laws of series and parallel combination

Circuit symbols and circuit diagrams.

While referring to a conductor or a semiconductor as an element possessing a resistance in an electrical or electronic (see chapter 19) circuit, one often uses the term ‘resistor’ (a term we have used earlier in this chapter). In other words, a resistor is a component used in a circuit, having a resistance that determines the current through it for any given voltage across its end-points in accordance with Ohm’s law (eq. (12-15)). Thus, a *resistor* is a component used in a circuit while *resistance* is an electrical property of the resistor. An electrical circuit consisting of various components like cells and resistors (other types of components will be mentioned later; of these, the capacitor has already been introduced in chapter 11) is conveniently represented with the help of certain *circuit symbols*. The circuit symbol of a resistor is shown in fig. 12-13(A), where the two dots represent the two *terminals* of the resistor, depicted with the zig-zag line. Fig. 12-13(B) shows the circuit symbol of a cell, in which the longer of the two contiguous line segments stands for the positive electrode and the shorter for the negative one. A simple circuit, made up of a cell and a resistor connected between the two terminals of the former is shown in fig. 12-13(C). The dots representing the terminals of a circuit component may sometimes be omitted in a circuit diagram.

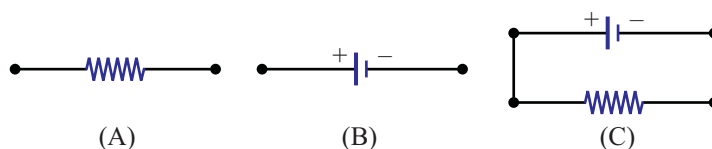


Figure 12-13: Use of symbols in representing electrical circuits; (A) a resistor; (B) a cell, or, more generally, a DC source of EMF (a DC source in brief; AC sources will be introduced in the next chapter); (C) a simple circuit made of a cell and a resistor; dots are used to denote the terminals of a circuit component, but may be omitted for the sake of simplicity.

Combinations of resistances.

In practical situations, more than one resistors are often connected in a circuit in a more or less complicated arrangement. The currents and voltages in any such arrangement can be analyzed in terms of two basic types of connection of the resistors: the *series* connection and the *parallel* connection. These two basic connections are shown in fig. 12-14(A), and 12-14(B) where, in each case, a combination of two resistances (R_1 , R_2) is connected between the two terminals of a cell of EMF E . Notice that, in a series combination, the same current flows through the resistors while the potential difference across these need not be equal, since the resistances need not be the same. On the other hand, in a parallel combination, the resistors have the same pair of terminals and so the potential differences across these have to be the same, while the currents flowing need not be the same since the resistances may differ.

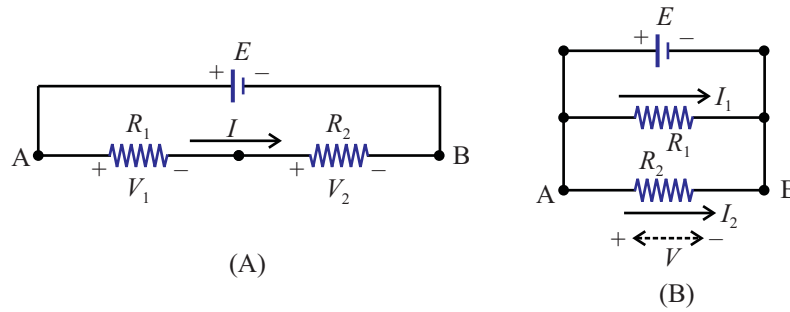


Figure 12-14: Series and parallel combination of resistances; (A) series combination of resistances R_1 and R_2 connected across a source of EMF E ; (B) parallel combination.

As I have mentioned earlier, the symbol E has been used to denote an EMF while, at places, the symbol \mathcal{E} has also been made use of. The symbol E is also used to denote the magnitude of electric field intensity. These two distinct uses of the symbol will not cause confusion since the meaning can be clearly read from the context.

Let me analyze the current-voltage relationships in fig. 12-14(A) in details. If the current passing through each of the resistances connected in series combination be I , then the potential differences across the two resistors will be, respectively,

$$V_1 = IR_1, \quad V_2 = IR_2, \quad (12-28)$$

and hence the potential difference across the *combination* (i.e., between the points A and B) will be

$$V = V_1 + V_2 = I(R_1 + R_2). \quad (12-29)$$

Now, imagine that, instead of the series combination of two resistances, a single resistance of such value (say, R) is connected across the same cell (as in fig. 12-13(C)) that the *same* current (I) flows through it as the one that flows through the combination. That resistance R is then termed the *equivalent resistance* of the series combination of R_1 and R_2 .

Evidently, since the same current is delivered by the cell in the two situations (one for the series combination and the other for the imagined equivalent resistance), the internal voltage drop will be the same in the two cases and hence, the same potential difference (sometimes referred to as ‘voltage’) V will appear across the equivalent resistance as between A and B in fig. 12-14(A). Indeed, the definition of equivalent resistance need not involve any reference to the cell or source of emf from which the current is drawn. The only condition that R has to satisfy in order to count as the equivalent resistance of R_1 and R_2 is that the same current (I) has to flow through it as through the series combination when the voltage drop across it is the same as that (V) between the two extreme terminals (A and B in fig. 12-14(A)) of the combination.

In other words, along with (12-29) we also have $V = IR$, and hence,

$$R = R_1 + R_2. \quad (12-30)$$

This tells us that *the equivalent resistance of two resistance R_1 and R_2 connected in series is simply the sum of the two*. I leave it to you to extend the above result for N ($N = 2, 3, \dots$) number of resistances in series, and to show that that

$$R = \sum_{m=1}^N R_m, \quad (12-31)$$

where the notation is self-explanatory.

One can analyze the current-voltage relations in fig. 12-14(B) in a similar manner. Let the current supplied by the cell (sometimes referred to as the ‘main current’) be I , while the currents through the two resistances be respectively I_1 , I_2 . Assuming that the potential difference between the two ends of the parallel combination (i.e., between the points A and B in fig. 12-14(B)) is V , we define the equivalent resistance of the combination as a resistance R which, imagined to be connected between A and B in place of the parallel combination, would result in the same current I being delivered by the cell and hence the same potential drop between A and B.

In the steady state there cannot occur any accumulation or decay of charge in any part of the circuit, and hence the current I delivered by the cell has to be equal to the sum $I_1 + I_2$ of the currents in the two branches of the parallel combination:

$$I = I_1 + I_2. \quad (12-32)$$

Reason this out. This is a particular instance of Kirchhoff’s first principle, explained in sec. 12.7.1 below.

Applying Ohm’s law to the two branches, we have:

$$V = I_1 R_1 = I_2 R_2. \quad (12-33)$$

Combining the two equations (12-32) and (12-33), we arrive at:

$$I = V \left(\frac{1}{R_1} + \frac{1}{R_2} \right). \quad (12-34)$$

Finally, applying Ohm’s law to the imagined situation in which the current I flowing through the equivalent resistance R results in a potential drop V across the latter, we get $I = \frac{V}{R}$, from which follows the formula relating the equivalent resistance of a parallel combination to the resistances of the two branches:

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2}. \quad (12-35a)$$

, i.e.,

$$R = \frac{R_1 R_2}{R_1 + R_2}. \quad (12-35b)$$

One can express this by saying that *the reciprocal of the equivalent resistance is the sum of the reciprocals of the resistances of the branches in the parallel combination*. Once again, this result (in the form (12-35a)) extends to the case of N number ($N = 2, 3, \dots$) of resistances joined in parallel, in which case, one has

$$\frac{1}{R} = \sum_{m=1}^N \frac{1}{R_m}, \quad (12-36)$$

where the notation is self-explanatory.

Problem 12-6

Two $10\ \Omega$ resistors are connected in parallel and the combination is then connected in series with a $20\ \Omega$ resistor and an ideal battery (i.e., one for which the internal resistance is negligibly small) of EMF $5\ \text{V}$. Calculate the current through the battery.

Answer to Problem 12-6

HINT: The parallel combination of two $10\ \Omega$ resistors can be replaced with a single equivalent resistor of resistance $5\ \Omega$. So, one now has to consider this $5\ \Omega$ resistor connected in series with a $20\ \Omega$ resistor and the $5\ \text{V}$ battery. Series connection means that the same current flows through all the three components. Since the internal resistance of the battery is negligible (ideal battery), the required current is $I = \frac{5}{5+20}\ \text{A}$, i.e., $0.2\ \text{A}$.

Problem 12-7

A $10\ \Omega$ resistance is joined in parallel with a $15\ \Omega$ resistance and the combination is joined to the terminals of a battery of EMF $3\ \text{V}$, when the potential difference across the combination of resistances is seen to be $2.4\ \text{V}$. If, now, the series combination of the two resistance be joined across the battery, what will be the potential difference between the terminals of the combination?

Answer to Problem 12-7

HINT: According to formula (12-35a), the equivalent resistance of the parallel combination is 6Ω , and hence, a potential difference of 2.4V across the combination corresponds to a current $\frac{2.4}{6}\text{A}$, i.e., 0.4A supplied by the battery. Since the internal drop in the battery is $(3 - 2.4) = 0.6\text{V}$, its internal resistance is $r = \frac{0.6}{0.4}\Omega$, i.e., 1.5Ω . For the series combination, the equivalent resistance is, according to formula (12-30), $R = 25\Omega$. The potential difference across the combination will now be, according to formula (12-20a), $V = \frac{R}{r+R}\mathcal{E}$, where $\mathcal{E} = 3\text{V}$. This gives $V = 2.83\text{V}$.

Problem 12-8

HINT: Two resistances, $R_1 = 6.0\Omega$ and $R_2 = 8.0\Omega$, both at $T = 20\text{K}$, are joined in parallel. The temperature coefficients of resistivity of the materials of the resistors are, respectively, $\alpha_1 = 40.0 \times 10^{-4}\text{K}^{-1}$ and $\alpha_2 = 20.0 \times 10^{-4}\text{K}^{-1}$. The temperature of the first resistor is raised by $\theta_1 = 2\text{K}$. By how much should the temperature of the second resistor be increased or decreased so that the equivalent resistance of the combination remains unchanged. What would your answer change if the two were connected in series?

Answer to Problem 12-8

HINT: Let the required change in temperature of the second resistor be θ_2 (positive in the case of an increase and negative for a decrease). Then, according to formula (12-17), the resistances get changed to $R'_1 = R_1(1 + \alpha_1\theta_1)$, and $R'_2 = R_2(1 + \alpha_2\theta_2)$ respectively. Noting that $\alpha_1\theta_1 \ll 1$ and assuming that, likewise, $\alpha_2\theta_2 \ll 1$ (this will be borne out by our results), we obtain (from the condition that the parallel combination should remain unchanged) $\frac{1}{R'_1} + \frac{1}{R'_2} = \frac{1}{R_1} + \frac{1}{R_2}$, where $\frac{1}{R'_i} \approx \frac{1}{R_i}(1 - \alpha_i\theta_i)$ ($i = 1, 2$). This gives $\theta_2 \approx -\frac{\alpha_1\theta_1 R_2}{\alpha_2 R_1}$. Using given values, we obtain $\theta_2 \approx -5.33\text{K}$, i.e., the temperature of the second resistor is to be decreased by nearly 5.3K . In the case of series combination, the condition that the equivalent resistance is to remain unchanged reads $R'_1 + R'_2 = R_1 + R_2$ which gives the result $\theta_2 = -\frac{\alpha_1\theta_1 R_1}{\alpha_2 R_2}$, i.e., the temperature of the second resistor is to be decreased by 3K .

12.6.2 Voltage division and current division

12.6.2.1 Voltage division

Fig. 12-15 depicts a set-up where a series combination of two resistances (R_1 and R_2) is connected across a source of EMF (E_0). A resistance X (or a combination of resistances) can be connected between the terminals A and B of the device. In the figure, the terminals of X are shown with two arrows, indicating that these can be connected to A and B if desired. If X is disconnected, the terminals A and B are said to be open-circuited, which is equivalent to assuming that a resistance $X \rightarrow \infty$ is connected between A and B. The internal resistance, if any, of the source of EMF E_0 is to be assumed to be included in R_1 .

One can now work out the potential difference between the terminals A and B in the open-circuit condition ($X \rightarrow \infty$) and also with a given resistance X connected between them. For the open-circuit condition, the voltage between A and B is

$$E = \frac{R_2}{R_1 + R_2} E_0. \quad (12-37)$$

Thus, the voltage appearing between A and B can be made equal to any chosen fraction of the EMF E_0 by an appropriate choice of R_1 and R_2 . One says that the voltage E_0 is *divided* in the ratio $R_1 : R_2$ across the two resistances R_1 and R_2 . A set-up for achieving voltage division is referred to as a *voltage divider*.

If, now, a resistance X is connected between A and B, the current through X and the potential difference across it (i.e., between the terminals A and B) are seen to be

$$I = \frac{E}{R' + X}, \quad E' = \frac{EX}{R' + X}, \quad (12-38a)$$

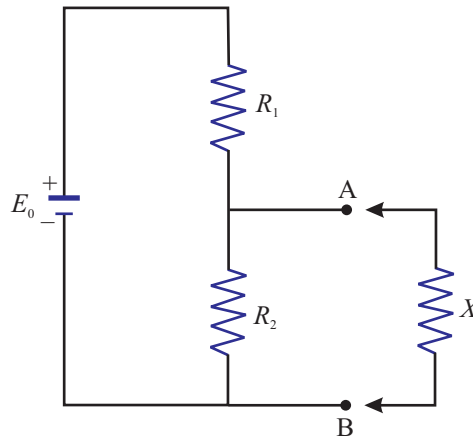


Figure 12-15: A voltage divider; a DC source of EMF E_0 is connected across the series combination of R_1 and R_2 ; the voltage appearing between A and B in the open circuit condition ($X \rightarrow \infty$) can be made equal to any chosen fraction of the EMF E_0 by an appropriate choice of R_1 and R_2 ; with respect to any resistance X that can be connected between A and B, the set-up acts as a DC source of EMF $E = \frac{R_2}{R_1 + R_2} E_0$, having an internal resistance $R' = \frac{R_1 R_2}{R_1 + R_2}$.

where E , R' are given by

$$E = \frac{E_0 R_2}{R_1 + R_2}, \quad R' = \frac{R_1 R_2}{R_1 + R_2}. \quad (12-38b)$$

In other words, the set-up shown in fig. 12-15 to the left of the terminals A and B (referred to as a voltage divider) acts as a source of EMF $E = \frac{R_2}{R_1 + R_2} E_0$, with an internal resistance R' , the latter being the parallel combination of the two resistances R_1 and R_2 .

A *voltage source* is a source of EMF E_0 with a *low* internal resistance (say, r), such that, when a resistance R is connected across its terminals, the voltage appearing across R (i.e., $\frac{E_0 R}{r + R} \approx E_0$) is independent of R .

Problem 12-9

Check out the result (12-38a).

Answer to Problem 12-9

HINT: The parallel combination of R_2 and X is, by (12-35b), equivalent to a single resistance $\frac{R_2 X}{R_2 + X}$.

Hence, when X is connected across R_2 , the current supplied by the source of EMF E_0 is $I_{\text{main}} =$

$\frac{E_0}{R_1 + \frac{R_2 X}{R_2 + X}}$, and the voltage appearing across the parallel combination is $E = \frac{I_{\text{main}} R_2 X}{R_2 + X}$. This, in turn, implies that the current through X is $I = \frac{I_{\text{main}} R_2}{R_2 + X} = \frac{\frac{E_0 R_2}{R_1 + R_2}}{X + \frac{R_1 R_2}{R_1 + R_2}}$. This verifies formula (12-38a), with E, R' given by (12-38b).

12.6.2.2 Current division

Fig. 12-16(A) shows a source of EMF E_0 in series with a *high* resistance R_0 . Such a device is termed a *current source* since the current through any resistance R connected between the terminals of the device is $I_0 = \frac{E_0}{R_0 + R} \approx \frac{E_0}{R_0}$, i.e., the current supplied by the device to the resistance R is *independent* of the value of R , as long as R is small compared to R_0 .

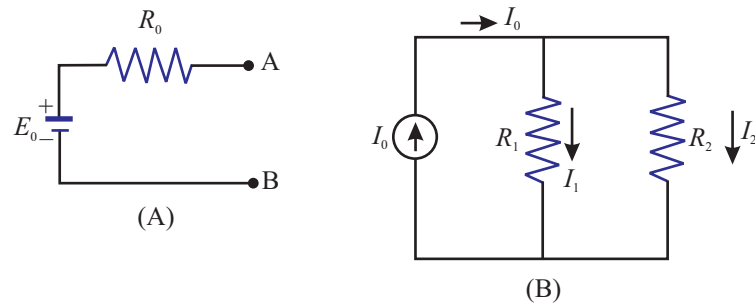


Figure 12-16: (A) A current source; (B) Current division.

Fig. 12-16(B) shows a current source I_0 (depicted by the symbol within the circle to the left of the figure; I_0 is referred to as the *short-circuit current* of the source) across which there is a parallel combination of two resistances R_1 and R_2 . The current I_0 sent out by the source is then divided into two currents I_1 and I_2 , where $I_0 = I_1 + I_2$ (refer to Kirchhoff's principle in section 12.7.1). The expressions for I_1 and I_2 are seen to be

$$I_1 = \frac{R_2}{R_1 + R_2} I_0, \quad I_2 = \frac{R_1}{R_1 + R_2} I_0. \quad (12-39)$$

In other words, the current I_0 sent out by the current source can be divided into two parts in any desired ratio by an appropriate choice of the resistances R_1 and R_2 . A set-up designed for such a purpose is termed a *current divider*.

If, now a resistance X (not shown in the figure) is connected between the two terminals of the set-up, then the latter can be seen to act as an equivalent current source whose short-circuit current and internal resistance can be adjusted to appropriately chosen values by choosing R_1 and R_2 , provided that X is small compared to R_1 , R_2 .

12.7 Analysis of DC electrical circuits

Figure 12-17 depicts an electrical circuit made up of a number of electrical cells and resistors. Any such circuit involves (a) a number of *branches* such as the branch AB or BC, where each branch includes one or more resistors and one or more electrical cells (or DC sources of EMF), (b) a number of junctions such as A or B where more than one branches meet, and (c) a number of closed loops such as ABCDA, each made up of more than one branches. At times, a circuit made up of more than one loops (or ‘meshes’) including resistances is referred to as a ‘network’.

The term DC (‘direct current’) circuit signifies a circuit carrying steady current in all its branches. The term may sometimes be used in the extended sense of a varying current flowing through a circuit or a branch thereof, but is to be contrasted with AC (‘alternating current’) where special arrangements are made use of to produce a current varying, typically, in a sinusoidal manner. Varying currents and AC will be our subject of discussion in chapter 13. A DC voltage source is a source with a steady EMF, while a DC *current* source i.e., a source supplying a steady current may also be included in a DC circuit. In what follows in the present chapter, the term ‘source of EMF’ or ‘voltage source’ will mean a DC source of EMF of negligibly small internal resistance, unless stated otherwise.

There exist a number of systematic procedures for the determination of the currents flowing through the various branches of such a circuit, and the corresponding potential drops across the resistors.

One fruitful approach is to make use of *Kirchhoff’s first and second principles*, while the

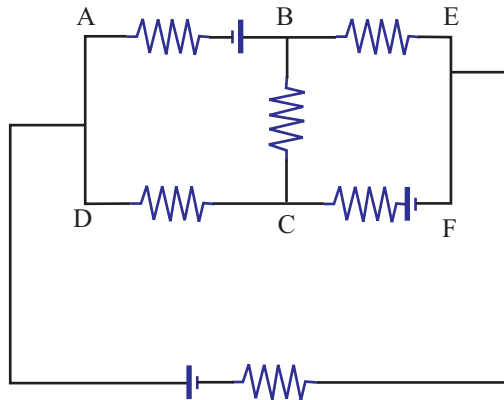


Figure 12-17: An electrical circuit including a number of branches, junctions, and closed loops.

principle of superposition (sec. 12.7.2) is of basic relevance. We assume that the currents in all the branches are steady DC currents.

12.7.1 Kirchhoff's principles

12.7.1.1 Kirchhoff's first principle

Consider any junction in an electrical circuit such as the junction A in fig. 12-18, in which only the branches meeting at the junction are shown. Let the currents through the various arms be as shown in the figure where, for the sake of convenience, all the currents are shown to flow *towards* the junction. In reality, some of the currents may flow *away* from the junction, in which case the numerical value of the corresponding current will carry a negative sign (a current of, say, -1A flowing towards a junction is equivalent to 1A flowing away from it).

Kirchhoff's first principle states that *the algebraic sum of all the currents flowing towards a junction in a circuit is zero*. Applied to the junction depicted in fig. 12-18, this means that

$$I_1 + I_2 + I_3 + I_4 = 0. \quad (12-40a)$$

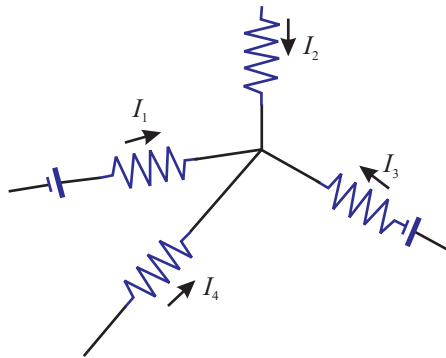


Figure 12-18: Four branches meeting at a junction A; illustrating Kirchhoff's first law, which requires that $I_1 + I_2 + I_3 + I_4 = 0$.

More generally, Kirchhoff's first principle can be expressed in the compact form

$$\sum I = 0, \quad (12-40b)$$

where the summation involves all the currents flowing towards the junction under consideration, with each of the currents carrying its appropriate sign.

This principle of Kirchhoff's is equivalent to the statement of the *principle of conservation of charge*. Since, in the steady state, charge cannot accumulate anywhere in the circuit, the net amount of charge converging to a junction in any given interval of time, and hence the net *rate* of flow of charge towards the junction, has to be zero.

12.7.1.2 Kirchhoff's second principle

Figure 12-19 depicts a closed loop (or *mesh*) ABCDA in an electrical circuit, where the other branches making up the circuit are not shown. The mesh includes a number of DC sources of EMF where the direction in which each source operates is determined by its polarities (shown with + and – symbols). The resistors in the various branches are marked in the figure, while the currents, in the respective *assumed* directions indicated by the arrows, are also shown. Finally, for the purpose of applying Kirchhoff's second law to the mesh, we choose a direction of traversal of the closed loop, say, the clockwise one.

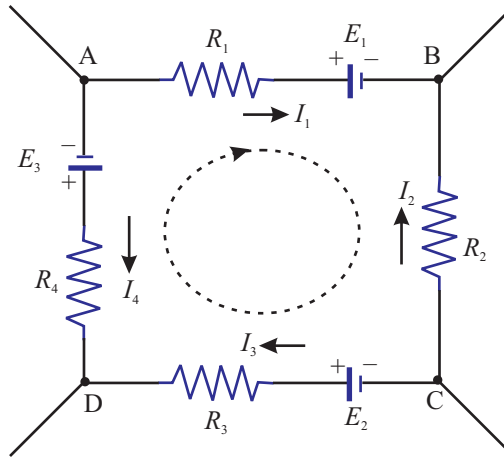


Figure 12-19: Illustrating Kirchhoff's second principle; the mesh ABCDA in a circuit (other branches and meshes of the circuit are not shown) includes a number of branches, each of which consists of a combination of resistances and sources of EMF, as the case may be; a certain sense of traversal of the mesh (say, clockwise, shown by the dotted line) is chosen; the application of the law then implies the relation (12-41a) for the mesh considered; the opposite choice for the sense of traversal would have led to the same relation.

Kirchhoff's second principle, applicable to any mesh in an electrical circuit states that *the algebraic sum of the EMFs operating in any chosen direction equals the algebraic sum of the potential drops across the resistors along the chosen direction.*

Applied to the mesh shown in fig. 12-19, Kirchhoff's second principle gives

$$I_1 R_1 - I_2 R_2 + I_3 R_3 - I_4 R_4 = -E_1 + E_2 - E_3, \quad (12-41a)$$

where the negative signs in the potential drops and the EMFs arise due to fact the assumed directions of the drops and the directions of operation of the EMFs are opposite to the chosen clockwise direction in the mesh. More generally, the law can be expressed in the compact form

$$\sum IR = \sum E, \quad (12-41b)$$

where the summations are to be performed with appropriate signs attached to the potential drops and the EMFs, with reference to the chosen direction (clockwise or anti-clockwise) of traversal of the mesh under consideration.

Incidentally, in applying Kirchhoff's second principle to any mesh involving one or more DC sources of EMF, the *internal resistances* of the sources have to be taken into account. Thus, in fig. 12-19, the internal resistances of the sources E_1 , E_2 , E_3 , have to be included in the resistances R_1 , R_3 , R_4 respectively. It is only then that the potential differences across their terminals can be taken to be the same as their respective EMFs. However, as mentioned earlier, the term 'source of EMF' (or, simply, 'source') is commonly used to refer to one with a negligibly small internal resistance.

This second principle of Kirchhoff's is equivalent to the statement that the electric field set up in a circuit carrying a steady current is *conservative* in nature (or, more precisely, conforms to the principle of conservation of energy). Thus, starting at any given point and going around a loop, the algebraic sum of the potential drops, i.e., the sum of the potential drops (including the drops across the double layers of the cells) taken with appropriate signs, has to be zero.

Considering the entire circuit, of which the mesh under consideration is a part, and taking into account the first principle, the second principle can be seen to imply that the total power delivered by the sources of EMF is the same as the power dissipated in all the resistances. This is nothing but the principle of conservation of energy for the circuit as a whole.

Kirchhoff's principles come in useful in the analysis of DC networks involving resistances and sources of EMF where, as mentioned above, a network may contain a number of interconnected meshes. These principles are applicable, with appropriate modifications, to the analysis of AC circuits as well (see sec. 13.5.4). Other useful rules for circuit analysis involve *Thevenin's* and *Norton's* theorems, and the *star-delta transformation*.

12.7.2 The principle of superposition

Imagine all the various resistances in any given electrical network to be labeled with integer indices running from 1 onward. Considering a resistor labeled k ($k = 1, 2, \dots$), let

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

the voltage across it be V_k and the current through it be I_k . These voltages and currents can be looked upon as the consequence of sources of EMF being included in the circuit. We will, however, not explicitly refer to these sources in our formulation of the principle of superposition.

The principle of superposition, and other principles of analysis of DC circuits are all applicable to AC circuits as well (refer to sec. 13.5.4) where the latter may contain circuit elements other than resistors, such as inductors and capacitors, and combinations thereof, characterized in terms of their *impedances*. The formulations of all these principles in the case of an AC circuit are analogous to those for a DC circuit. For the time being, however, we will confine ourselves to the analysis of DC circuits, where only DC sources and resistors will be considered.

For the given network, as long as the sources of EMF are fixed, the set of voltages (V_1, V_2, \dots) (briefly denoted by $\{V_k\}$) and also the corresponding set of currents ($I_1, I_2, \dots; \{I_k\}$ in brief) remain fixed.

Consider now the following two situations. In one of these the sets of voltages and currents are, respectively, $\{V'_k\}$ and $\{I'_k\}$ while, in the other, these are $\{V''_k\}$ and $\{I''_k\}$, the change being brought about by changes in the sources of EMF which, however, need to be explicitly specified in the present context.

If, now, the sources of EMF be such that the set of voltages is modified to $\{V'_k + V''_k\}$ then, the principle of superposition states that the set of currents will get modified to $\{I'_k + I''_k\}$.

The principle of superposition is a consequence of the relation of linearity between the voltage across a resistor and the current flowing through it.

There exist a class of circuit elements, mostly used in *electronic* circuits (refer to chapter 19) for which the currents and voltages are related in a *non-linear* manner. The principle of superposition does not apply to circuits involving such nonlinear elements.

Here is an alternative formulation of the principle of superposition, essentially identical to the one stated above.

Let the set of EMF's of the DC sources in the given network be (E_1, E_2, \dots) . Suppose that, for values $(E_1 = E'_1, E_2 = E'_2, \dots)$ of the EMF's in this set, the voltages across the resistances R_k ($k = 1, 2, \dots$) be V'_k , and the currents through these be I'_k , and again, for values $(E_1 = E''_1, E_2 = E''_2, \dots)$ of the EMF's, the corresponding voltages and currents be V''_k, I''_k , where the change in the EMF's is supposed to be effected *without any change in the internal resistances of these sources*. If, now, the values of the EMF's be $(E_1 = E'_1 + E''_1, E_2 = E'_2 + E''_2, \dots)$, then the voltages and currents will get changed to $V'_k + V''_k, I'_k + I''_k$ ($k = 1, 2, \dots$) where, once again, the internal resistances of the sources of EMF are assumed to remain unchanged.

The constraint requiring that the internal resistances of the sources are to remain unchanged may appear to be an unduly restrictive one. In practice, sources of EMF are commonly close to being *ideal* ones in that their internal resistances are negligibly small. If, on the other hand, the internal resistances cannot be assumed to remain unchanged, then alternative version of the principle stated earlier, where the sources of EMF are not explicitly referred to, is to be made use of.

12.7.3 The Wheatstone bridge

Fig. 12-20 shows a simple electrical network referred to as the Wheatstone bridge, consisting of four resistances (one or more of which may be combinations of several resistances; the figure depicts equivalent resistances of such combinations, if present). A DC source of EMF is connected between terminals marked 'A' and 'C', while a an ammeter (E; a current measuring device) is connected between 'B' and 'D'. The resistances R_1, R_2, R_3, R_4 in the four *arms* AB, BC, AD, DC are so adjusted that the ammeter shows no deflection. This is referred to as the *null* condition (or *balanced* condition) of the bridge. In this condition the current through the arm BC is the same as that through AB (denoted by I_1 in the figure) by Kirchhoff's first principle while, likewise, the current through DC is the same as that through AD (I_2). At the same time, the potential

difference between B and D has to be zero, which gives

$$I_1 R_1 = I_2 R_3, \quad I_1 R_2 = I_2 R_4, \quad (12-42a)$$

i.e.,

$$\frac{R_1}{R_3} = \frac{R_2}{R_4}. \quad (12-42b)$$

Thus, knowing any three of the four resistances in the arms of the Wheatstone bridge, one can determine the fourth.

The Wheatstone bridge is a common device employed for the determination of resistances.

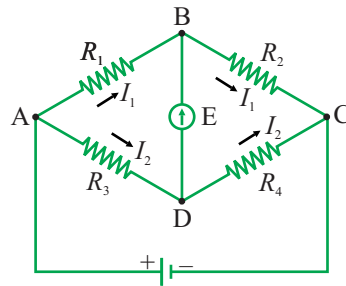


Figure 12-20: The Wheatstone bridge; AB, BC, AD, DC are the four arms of the bridge, carrying resistances R_1, R_2, R_3, R_4 , which are related as in (12-42b) when the bridge is in the balanced condition, i.e., when the current through the ammeter E is zero.

Problem 12-10

Consider the circuit shown in fig. 12-21(A) in which four equal resistances $R = 18 \, \Omega$ are connected to form a closed loop ABCD, with a resistance $R' = 10 \, \Omega$ connected between the junctions B and D. Current is driven by a battery of EMF $E = 4 \, \text{V}$ and internal resistance $r = 2 \, \Omega$. Find the current I sent out by the battery and the power dissipated in each of the resistances R as also in the battery itself.

Answer to Problem 12-10

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

HINT: From the symmetry of the arrangement, the potentials at B and D have to be equal, which means that there is no current in the resistance R' (you can check this out by working longhand and setting up equations by making use of Kirchhoff's principles, and then solving for the currents in the various branches of the circuit from these equations; R' may as well be omitted altogether from the circuit arrangement without affecting the current in the other branches). Thus the circuit effectively consists of a parallel combination of two resistances, each of which is a series combination of two resistances R . This gives, finally, an equivalent resistance R between the points A and C, i.e., in other words, the current I sent out by the battery is the same as that in the circuit of fig. 12-21(B). Using given values, one gets $I = \frac{E}{r+R} = 0.2$ A. This current is divided into two equal parts at the junction A, i.e., the current through each of the four resistances R is $I' = 0.1$ A, and the rate of energy dissipation in each is $W = I'^2 R = 0.18$ W. The power dissipated in the battery is $W' = I^2 r = 0.08$ W.

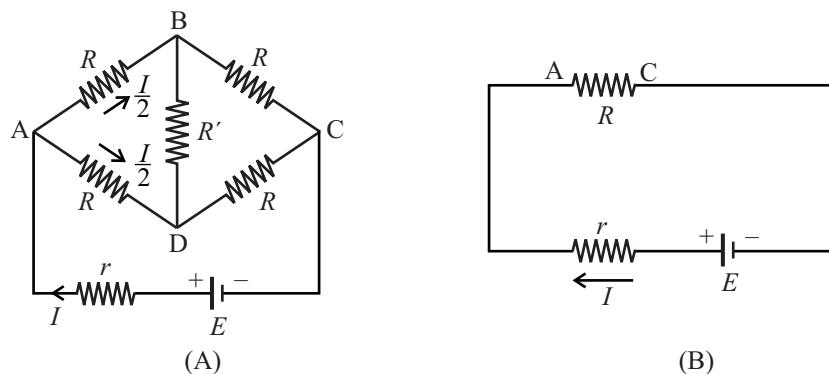


Figure 12-21: (A) A circuit arrangement showing four identical resistances R forming a loop ABCD, with a battery of internal resistance r and EMF E connected to the combination; because of the symmetry of the arrangement, the resistance R' , connected between B and D, does not carry any current, and one can then determine the current I sent out by the battery by looking at the equivalent circuit arrangement (B), where the equivalent resistance between the points A and C is R .

Problem 12-11

Suppose you are given a number of $10\ \Omega$ resistances, each capable of dissipating a maximum 1 W of power. What combination of these, using the minimum possible number of resistors, will have an equivalent resistance of $10\ \Omega$ and will be capable of dissipating 25W?

Answer to Problem 12-11

HINT: Evidently, one needs to combine the resistors both in series and parallel, since otherwise the equivalent resistance would be either greater than or less than 10Ω . Moreover, if one combines n number of resistors in series then one would need to connect in parallel n number of such combinations so as to make up an equivalent resistance of 10Ω . This requires, in all, n^2 number of resistors. Assuming that the maximum possible power (i.e., $1W$) is made to be dissipated in each resistor, the minimum number of resistors required for a total dissipation of $25W$ is seen to be 25. This requires 5 branches to be connected in parallel, where each branch is a series combination of 5 resistors (alternatively, 5 resistors connected in parallel, and 5 such connected in series).

Note that, considered in isolation, each resistor requires a supply of $V = \sqrt{10}V$ for the maximum possible power dissipation ($1W$) to occur. The combination, on the other hand, requires a supply of $V' = 5\sqrt{10}V$ for the maximum possible power dissipation ($25W$).

Problem 12-12

Twelve resistors, each of resistance $R = 1\Omega$ are connected as in fig. 12-22, forming a network with eight junctions and twelve branches, where three branches meet at each junction. Find the equivalent resistance of the network between junctions marked A and G.

Answer to Problem 12-12

SOLUTION: It helps to visualize the network in the form of a cube ABCDEFGH as in fig. 12-22. Though the cubical form is not of direct relevance in solving the problem (only the connections of the resistors between the junctions is relevant), the symmetry of the cube helps in identifying a number of symmetry relations between the currents in the various branches of the network.

The dotted lines in the figure indicate the leads through which the main current (I) from the source (not shown) of EMF flow into and out of the network at junctions A, G.

If V be the potential difference between A and G, then the equivalent resistance is given by $R_{eq} = \frac{V}{I}$. The main current I breaks up into three equal parts (by symmetry of the resistances of the network; instead of invoking this symmetry, one can work longhand, considering one by one the independent meshes and junctions, and making use of Kirchhoff's principles; symmetry

considerations make the solution simpler), each $\frac{I}{3}$, flowing along branches AB, AD, AE. At E, the current $\frac{I}{3}$ along AE breaks up into two equal parts, each $\frac{I}{6}$, along EF, EH (again, by symmetry of connection of the resistors, all of which are equal). The currents along FG, HG, CG must again be all equal, i.e., $\frac{I}{3}$, flowing towards G and making up the main current I flowing out from G (in other words, we invoke Kirchhoff's first principle at junctions A, E, G, along with the symmetry of the network connections).

Thus, the potential difference between A and G is, in a notation that is self-explanatory, $V_{AG} = V_{AE} + V_{EF} + V_{FG} = (\frac{I}{3} + \frac{I}{6} + \frac{I}{3})R = \frac{5IR}{6}$. This implies that the required equivalent resistance between A and G is (with $R = 1\Omega$,) $R_{eq} = \frac{5}{6}\Omega$.

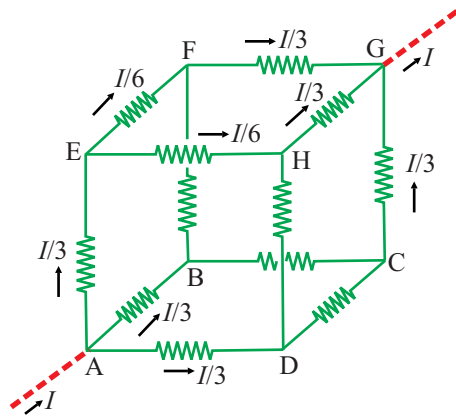


Figure 12-22: A network of twelve identical resistors, each of resistance R ; the network has eight junctions and twelve branches, with three branches meeting at each junction, and can be visualized as a cubic frame ABCDEFGH; the geometrical symmetry of the cube helps identify a number of symmetry relations between the currents in the various branches, where these symmetry relations are solely the consequence of all the resistors being identical and of the way they are connected; the dotted lines indicate main current from a source (not shown), entering and leaving the network at junctions A and G respectively; the equivalent resistance between A and G, as worked out in problem 12-12, is $\frac{5}{6}R$.

Problem 12-13

The network shown in fig. 12-23(A) includes two sources P, Q, of which P has an EMF $E_1 = 1.5V$, and an internal resistance $r_1 = 0.5\Omega$, the corresponding values for Q being $E_2 = 3V$, and $r_2 = 1\Omega$, where the polarities of the sources are as shown. With resistances $R_1 = 1.5\Omega$, $R_2 = 1\Omega$, $R_3 = 2\Omega$ connected as in the figure, obtain the current through R_3 , first by invoking the principle of

superposition, and then by making use of Kircchoff's principles.

Answer to Problem 12-13

HINT: The circuit diagram of fig. 12-23(A) is redrawn in fig. 12-23(B), with the internal resistances r_1, r_2 added to R_1, R_2 so as give resistances $R'_1 = 2\Omega, R'_2 = 2\Omega$, as a result of which P and Q can now be looked upon as ideal sources (having zero internal resistance) with EMF's $E_1 = 1.5V$, $E_2 = 3V$, and with polarities as indicated with the '+' and '-' symbols.

Recalling the principle of superposition, we now consider two situations, of which the first corresponds to *short-circuiting* the source P (now considered to be an ideal one) and the second to short-circuiting Q (again, an ideal source now). This implies that, in the first situation, we have $E_1 = E'_1 = 0, E_2 = E'_2 = 3V$, while in the second situation, the corresponding values are $E_1 = E''_1 = 1.5V, E_2 = E''_2 = 0$, the resistances $R'_1 = 2\Omega, R'_2 = 2\Omega, R_3 = 2\Omega$ being the same in the two cases (draw your own circuit diagrams representing these two situations; note that what this effectively means is that we are imagining the EMF's of the sources to be changed while keeping the internal resistances intact).

Considering now the first situation, we effectively have an ideal source of EMF $E'_2 = 3V$ supplying current to the series combination of $R'_2 = 2\Omega$ and $R' = 1\Omega$, the latter being the equivalent resistance of the parallel combination of $R'_1 = 2\Omega$ and $R_3 = 2\Omega$. The main current is then $I' = 1A$, flowing from the '+' to the '-' terminal of Q. This current is divided equally at junction B shown in fig. 12-13(A) between R'_1 and R_3 (reason out why), so that the current through R_3 is $I'_3 = 0.5A$ flowing from junction B to A.

Similar considerations, now for the second situation indicated above, gives the current 0.25A, through R_3 , flowing from A to B, i.e., to a current $I''_3 = -0.25A$ flowing from B to A. Thus, by the superposition principle, the current through R_3 for the situation depicted in fig. 12-23(A) is $I_3 = I'_3 + I''_3 = 0.25A$ in the direction from B to A.

We now look at another approach for solving the problem, making use of Kirchhoff's principles. For this we refer to fig. 12-23(B), with currents I_1, I_3 through R'_1 and R_3 in assumed directions indicated by arrows, and with current $I_2 = I_1 + I_3$ through R'_2 , again in the direction of the arrow, as required by Kirchhoff's first principle.

Looking at the mesh ABCD, and considering the assumed direction of circulation shown by the

bent dotted arrow, Kirchhoff's second principle gives $I_1 R'_1 - I_3 R_3 = E_1$, and similarly, for the mesh ABEF, $I_3 R_3 + (I_1 + I_3) R'_2 = E_2$. Using values $R'_1 = R'_2 = R_3 = 2\Omega$, and $E_1 = 1.5\text{V}$, $E_2 = 3\text{V}$, one obtains, once again, $I_3 = 0.25\text{A}$, flowing in the direction from junction B to junction A.

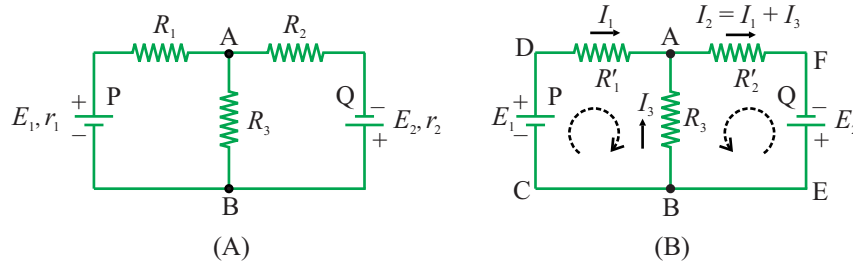


Figure 12-23: (A) An electrical network made of sources of EMF P,Q, with EMF's E_1 , E_2 and internal resistances r_1 , r_2 , and of resistances R_1 , R_2 , R_3 , as mentioned in problem 12-13 (B) The same network as in (A), but with the internal resistances of P, Q added to R_1 , R_2 respectively, giving new resistances R'_1 , R'_2 , while the sources P, Q are now ideal ones, with zero internal resistance; dotted bent arrows indicate assumed directions of circulation in the meshes ABCD and ABEF, for invoking Kirchhoff's second principle; the currents I_1 , I_3 , and $I_2 = I_1 + I_3$ in the directions of the arrows conform to Kirchhoff's first principle; the value of I_3 (as also of any of the other currents in the circuit) can be obtained either by invoking the principle of superposition or by making use of Kirchhoff's principles, as in the problem.

Problem 12-14

Identical resistors, each of resistance R , are connected to form an infinitely extended network as shown in fig. 12-24, where the network can be visualized as a planar array of junctions, with four resistors meeting at each junction, and the the junctions forming an infinitely extended square lattice; (A) imagining a source to be connected between the points A (an arbitrarily chosen junction in the network) and B (an adjacent junction) marked in the figure, find the equivalent resistance between A and B; (B) find the equivalent resistance between A and C (a junction diagonally opposite to A).

Answer to Problem 12-14

Considering the array of junctions, which can be visualized as a planar square lattice, one can assign a pair of integer co-ordinates to each junction, starting with the point A in fig. 12-24, an arbitrarily chosen junction, as $(0, 0)$, where the first co-ordinate increases in the horizontal direction towards the right of the figure and the second co-ordinate increases upward along the vertical direction. Thus, the junctions B and C have co-ordinates $(1, 0)$ and $(1, 1)$.

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

We solve the problem by invoking the superposition principle. Thus, imagine a situation in which only one lead carries a main current I into the network at the junction A i.e., at $(0, 0)$, and gets divided as it flows outward to infinitely distant junctions, where the currents in the branches become infinitesimally small. We can then imagine that these currents, each tending to zero, are picked up by returning leads at infinitely distant junctions that join into one single lead to make up the main returning current I . Let V_{mn} be the potential at the junction with co-ordinates (m, n) ($m, n = 0, \pm 1, \pm 2, \dots$) (we will henceforth refer to it as the 'junction (m, n) ' for the sake of brevity) with respect to infinitely remote junctions ('junctions at infinity'). By the symmetry of the network, since junctions $(\pm m, \pm n)$ are symmetrically situated with respect to $(0, 0)$, the V_{mn} 's satisfy the relations

$$V_{mn} = V_{\pm m, \pm n}. \quad (\text{A})$$

Again, imagine the complementary situation where vanishingly small currents enter into the network at all the junctions at infinity and are eventually picked up by a single returning lead at any point, say at the junction (p, q) ($p, q = 0, \pm 1, \pm 2, \dots$). Since all points in the array are equivalent in the absence of leads, the potential at junction $(p + m, q + n)$ will now be $-V_{m, n}$ (reason out why). By the principle of superposition, if a current I is injected into the network at $(0, 0)$ and returned to the source at (m, n) , then the difference of voltages at $(0, 0)$ and (m, n) will be $2(V_{00} - V_{mn})$ (check this out).

(A) The case of main current I entering at A $((0, 0))$ and leaving at B $((1, 0))$ can be worked out directly, without evaluating the general expression for V_{mn} . Considering the situation where the main current enters at A and then gets divided indefinitely down to infinitely remote points, one observes that the current is divided into four equal parts at A, and hence the current from A to B is $\frac{I}{4}$, which gives $V_{00} - V_{10} = \frac{IR}{4}$. Thus the required equivalent resistance in this case, obtained by invoking the superposition principle as explained above, is $\frac{R}{2}$.

(B) The case where the main current I enters at A and leaves at C $((1, 1))$ is, by comparison, quite non-trivial and can be worked out by first evaluating V_{mn} , for arbitrarily chosen values of m, n , in the situation where the current enters at A and returns through infinitely remote leads as the summation of infinitesimally small currents.

Imagine, in this situation, the currents from the junction (m, n) directed towards junctions $(m, n + 1)$, $(m + 1, n)$, $(m - 1, n)$, $(m, n - 1)$. Applying Kirchhoff's first principle to these currents flowing

out from the junction (m, n) , one obtains the following basic formula for the network:

$$\frac{1}{R}(V_{m,n+1} + V_{m,n-1} + V_{m+1,n} + V_{m-1,n} - 4V_{m,n}) = 0. \quad (B)$$

Along with the symmetry relations $V_{mn} = V_{\pm m, \pm n}$ mentioned above, the V_{mn} 's satisfy the additional symmetry relation

$$V_{m,n} = V_{n,m} \quad (m, n = 0, \pm 1, \pm 2, \dots). \quad (C)$$

(reason this out).

We now make the *ansatz*

$$V_{m,m+k} = V_0 \lambda^m \mu^k \quad (\text{for } m, k \geq 0), \quad (D)$$

where λ, μ are to be determined from the basic relation (B) and the value of V_{mn} for general values of the indices m, n are to be obtained from the symmetry relations (A) and (C). Because of the restrictions on m, k in (D), the following boundary conditions are to be satisfied in addition to (D):

$$4V_{0k} = V_{0,k+1} + V_{0,k-1} + 2V_{1,k}, \quad (V_{00} - V_{01}) = \frac{IR}{4}. \quad (E)$$

(boundary condition for junctions with $m = 0, k > 0$, and for the junction with $m = 0, k = 0$). Substituting (D) in (B), (E), we obtain $\mu = 3 - 2\sqrt{2}$, $\lambda = -3 + 2\sqrt{2}$ (i.e., $V_{m,m+k} = (-1)^m V_0 \mu^{m+k}$ ($m, k \geq 0$), where V_0 is obtained below). Then, making use of these results in (D) for the special values $m = 0, k = 0$ and $m = 0, k = 1$, and invoking the second relation in (E), we obtain $(V_{00} - V_{01}) = V_0(1 - \mu) = V_0(2\sqrt{2} - 2) = \frac{IR}{4}$ and, again, with $m = 1, k = 0$, $V_{11} = V_0 \lambda = V_0(-3 + 2\sqrt{2})$. The equivalent resistance between A and C then comes out from the defining formula $IR_{AC} = 2(V_{00} - V_{11})$ (in accordance with the principle of superposition as explained above) as $R_{AC} = \frac{1}{\sqrt{2}}R$.

We can also check that the equivalent resistance between A and B, defined by $IR_{AB} = 2(V_0 - V_{01})$, evaluates to $R_{AB} = \frac{1}{2}R$, in conformity with the result obtained above by more direct reasoning.

12.8 The magnetic effect of currents

Electrostatics starts from Coulomb's law, which is derived from forces observed between charged bodies. An analogous force of interaction is found to operate between another set of bodies known as *magnets*. Magnets exert forces not only on one another, but on

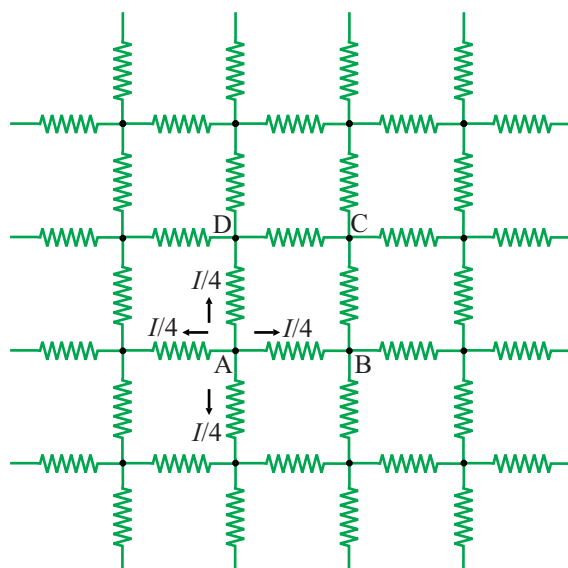


Figure 12-24: An infinitely extended network of identical resistors, each of resistance R ; it is convenient to visualize the network as an infinite square lattice of junctions, with four branches meeting at each junction; main current from a source can be made to flow into and out of the network by means of leads (source and leads not shown); with leads connected between points A and B, the equivalent resistance of the network turns out to be $\frac{R}{2}$, while the equivalent resistance between A and C works out to $\frac{R}{\sqrt{2}}$, as in problem 12-14.

other bodies made of certain special materials, known as magnetic materials (forces on bodies made of other types of materials are, in general, negligibly small). Moreover, a body made of a magnetic material, even though not a magnet to start with, can be made to become one with the help of other magnets or by means of an *electrical current*. What is more, an electrical current is found to exert force on a magnet. And finally, *currents are found to exert forces on one another*. This last observation can be made use of to explain, from a fundamental point of view, almost all observed phenomena relating to magnets, magnetic materials, and the forces of interaction between them.

12.8.1 Force between currents composed from elementary forces

Consider two current-carrying loops of thin wire brought close together as in fig. 12-25(A). It is found experimentally that a force is exerted by any one of the loops on the other and that this force is proportional to the product of the currents flowing in the two loops. It also depends on the shapes of the loops and other geometrical factors relating to their position and orientation with respect to each other. Indeed, it is this observed

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

force that is made use of in defining the *unit of current* (the ampere, A) in the SI system. By means of a process of analytical reasoning based on experimental observations, it has been found that a formula for the force between two current-carrying wires consistent with all these experimental observations can be obtained by assuming that this force is the vector sum of a large number of small ‘elementary’ forces.

To understand what this means, consider a small elementary portion of length in each loop, shown in a magnified view in fig. 12-25(B). These can be treated as small straight lines of lengths, say δl_1 , δl_2 , giving us two vectors $\delta l_1 \hat{n}_1$, $\delta l_2 \hat{n}_2$, where \hat{n}_1 , \hat{n}_2 are unit vectors along the lengths of the two elements of length, pointing in the directions of the currents flowing in the respective wires. These two vectors we call the vector length elements in the two wires. One can now associate a force of small magnitude (say, δF) with this pair of vector length elements and call it the force exerted by one element on the other. To be more specific regarding *which* element exerts what force on the other, we can use the symbol, say, δF_{12} to denote the force *on* the first element exerted *by* the second. The force on the second element exerted by the first would then be denoted by δF_{21} , these two being related to each other as $\delta F_{12} = -\delta F_{21}$ (Newton’s third law).

You should be careful how you read an expression like δl or δF , as also similar expressions elsewhere in the book. For instance, δl is not δ times l - it denotes one single quantity in itself, namely a *small* length. At times, when l stands for a length by itself, the symbol δl is used to mean a small *increment* in l . Similarly, δF stands for a force of small magnitude or, depending on the context, a small increment in the force F . One often uses symbols like dl or dF in place of δl or δF while referring to these quantities *in the limit* of the relevant magnitudes going to zero. These are then termed *infinitesimal* quantities.

By considering all possible pairs of such length elements and calculating all the corresponding elementary forces, the expression for which I am going to give you below, one can then take the vector sum of these elementary forces to arrive at the force exerted

by, say, the second wire on the first. One can express this symbolically as

$$\mathbf{F}_{12} = \sum \delta \mathbf{F}_{12}, \quad (12-43)$$

where the summation is to be carried out over all possible pairs of length elements in the two wires. A similar expression would give you \mathbf{F}_{21} , which would be related to \mathbf{F}_{12} as

$$\mathbf{F}_{21} = -\mathbf{F}_{12}. \quad (12-44)$$

This is referred to as the *magnetic force* between the two current carrying wires.

Thus the magnetic force between two current-carrying wires (these wires are usually in the form of closed loops or form parts of closed circuits; the loops or circuits include sources of EMF driving currents through the wires; the latter have not been shown in fig. 12-25) can be worked out by referring to the elementary forces between pairs of length elements in these wires, and everything is thus known in principle once one has the expression for the typical elementary force between a typical pair of length elements. In this sense the situation is similar to the force between two charged bodies which can be worked out in principle by summing over the elementary forces between pairs of point charges (or pairs of small volume elements of charge in the two bodies), the typical elementary force in question being given by Coulomb's law (sec. 11.3.1).

The corresponding expression for the elementary magnetic force between two length elements in a pair of current-carrying loops does not look a particularly neat one, but let me give it to you nevertheless, referring for the sake of simplicity to a situation where the two current loops are assumed to be situated in vacuum:

$$\delta \mathbf{F}_{12} = \frac{\mu_0 I_1 I_2}{4\pi} \frac{(\delta l_1 \hat{n}_1) \times ((\delta l_2 \hat{n}_2) \times \mathbf{r}_{12})}{r_{12}^3}. \quad (12-45)$$

In this formula, μ_0 is a constant termed the *permeability of free space*, its value being $4\pi \times 10^{-7} \text{ N}\cdot\text{A}^{-2}$. The formula gives the force on a length element δl_1 situated at the point, say, \mathbf{r}_1 on the first wire exerted by another length element δl_2 at, say, \mathbf{r}_2 on the second wire, where the unit vectors \hat{n}_1 and \hat{n}_2 have been defined above. The vector \mathbf{r}_{12} denotes

the vector distance extending from \mathbf{r}_2 to \mathbf{r}_1 , with magnitude $r_{12} \equiv |\mathbf{r}_{12}|$.

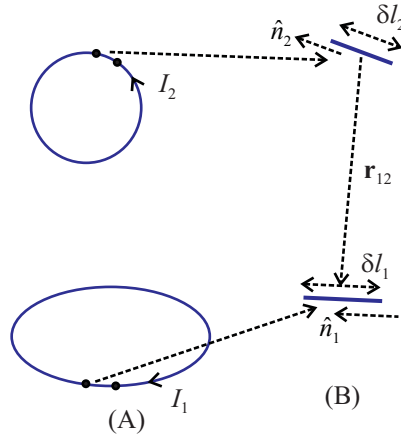


Figure 12-25: Magnetic force between two current-carrying loops of wire; (A) two loops of wire, each carrying a current and consequently exerting a force on the other; each loop can be imagined to be made up of a large number of small length elements, a pair of such elements being shown magnified in (B); the force between this pair is given by eq. (12-45) and by Newton's third law; the force between the two loops is then obtained by summing these elementary forces (equations (12-43) and (12-44)); the sources of EMF driving the currents in the loops are not shown.

Problem 12-15

Calculate the force of interaction between two small elements of current-carrying loops of wire, where one of the elements, of length $\delta l_1 = 1.0 \times 10^{-3}\text{m}$ is placed at the origin (O) of a right-handed Cartesian co-ordinate system along the x-axis, with a current $I_1 = 1.5\text{A}$ pointing along the unit vector $\hat{n}_1 = \hat{i}$, while the other element, of length $\delta l_2 = 2.0 \times 10^{-3}\text{m}$ is placed at the point P with co-ordinates $(1, 1, 1)\text{m}$, with a current $I_2 = 1.0\text{A}$, the unit vector along the direction of flow of current being $\hat{n}_2 = \frac{1}{\sqrt{2}}(\hat{j} + \hat{k})$ ($\hat{i}, \hat{j}, \hat{k}$ are the unit vectors along the three co-ordinate axes).

Answer to Problem 12-15

SOLUTION: We make use of formula 12-45 with $\mathbf{r}_{12} = -(\hat{i} + \hat{j} + \hat{k})$ (vector distance from the second element to the first element), $\hat{n}_1 = \hat{i}$, $\hat{n}_2 = \frac{1}{\sqrt{2}}(\hat{j} + \hat{k})$, so as to obtain $\delta \mathbf{F}_{12} = \frac{\mu_0}{4\pi r_{12}^3} I_1 I_2 \delta l_1 \delta l_2 \hat{n}_1 \times (\hat{n}_2 \times \mathbf{r}_{12})$. Here $\hat{n}_1 \times (\hat{n}_2 \times \mathbf{r}_{12}) = (\hat{n}_1 \cdot \mathbf{r}_{12})\hat{n}_2 - (\hat{n}_1 \cdot \hat{n}_2)\mathbf{r}_{12} = -\hat{n}_2$; hence, substituting values, $\delta \mathbf{F}_{12} = -4.1 \times 10^{-14}(\hat{i} + \hat{j})\text{N}$; the magnitude of the force of interaction is $5.8 \times 10^{-14}\text{N}$.

12.8.2 Force between a pair of parallel current-carrying wires

The formula (12-45), then, can be looked at as the starting point in the *magnetic effect of steady currents*, similar to Coulomb's law of force between two point charges or two small volume elements of charge. It looks complicated owing to a number of vectors and vector cross products appearing in it, and things don't get any more easy when one tries to sum over the elementary forces to work out the force between two current-carrying wires. In some cases, however, the end result appears simple. As an illustration, I give below the formula for the force *per unit length* felt by the first wire carrying current I_1 due to a second wire carrying current I_2 , the two wires being both long and parallel to each other. The formula looks simple and is, at the same time, instructive:

$$\mathbf{F}_{12} = \frac{\mu_0}{2\pi} \frac{I_1 I_2}{d} \hat{n}. \quad (12-46)$$

In (12-46) d denotes the distance between the two parallel wires, \hat{n} is a unit vector lying in the plane of the wires but perpendicular to both of those, directed from the first to the second wire (fig. 12-26), and the product of the currents, $I_1 I_2$, is to be taken with appropriate sign, being negative if the two currents are in opposite directions. While (12-46) gives the force *on* the first wire exerted *by* the second, the force on the second wire exerted by the first would simply be $\mathbf{F}_{21} = -\mathbf{F}_{12}$.

The formula tells us that the force between two current-carrying wires is *an attractive or a repulsive one if the currents are in the same or opposite directions respectively*. The magnitude of this force can be made use of to define operationally the unit of current (A) as that current which, flowing through each of two long parallel wires 1 m apart causes a force of 2×10^{-7} N to be exerted on unit length of either wire. In reality, however, the operational definition of the ampere in the SI system is made in terms of the force between a pair of *circular* coils.

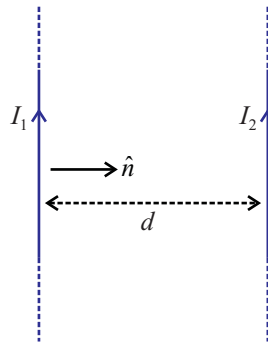


Figure 12-26: Magnetic force between two infinitely long, straight, and parallel current-carrying wires; a portion of each wire is shown; each of the wires is assumed to be a closed circuit, the remaining parts of which are assumed to be at infinitely large distances; the currents I_1 and I_2 are to be taken with appropriate signs - assuming any one of the two possible directions (upward and downward in the figure) as positive, a current flowing in the opposite direction is to be assigned a negative sign; the force is given by formula (12-46), and by Newton's third law.

12.8.3 Magnetic field intensity

While equation (12-45) can be looked at as the basis of the magnetic effect of steady currents and, in the ultimate analysis, of magnetism as such, a more convenient concept to work with is that of the *magnetic field* and *magnetic field intensity*, similar to the approach adopted in electrostatics.

The concepts of electric and magnetic fields and their intensities are, however, not just convenient theoretical constructs; the electric and magnetic fields actually constitute a *dynamical system* distributed throughout space.

The fact that a current-carrying wire exerts force on other current-carrying wires and on magnets in its vicinity, tells us that a current loop creates a certain region of influence around itself, an influence that causes a force to be exerted on other currents and magnets. This idea of an influence set up around a current can be made quantitative by means of the concept of *magnetic field intensity* at any given point in space which is defined in terms of the force on a current element placed at that point.

While the force is exerted on closed current loops (or on wires forming electrical circuits) it can be looked upon as the vector sum of elementary forces on small length elements

of the loops. Thus, consider a length element δl at the point with position vector \mathbf{r} in a current loop carrying a current I , the unit vector along the element in the direction of flow of the current being, say, \hat{n} (fig. 12-27). Then the magnetic intensity at \mathbf{r} is defined as a vector (say, \mathbf{B}) such that the elementary force exerted on the length element is given by

$$\delta \mathbf{F} = I \delta l \hat{n} \times \mathbf{B}. \quad (12-47)$$

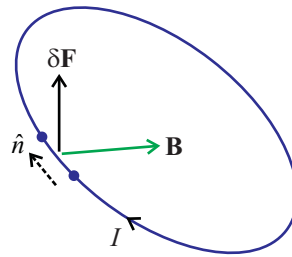


Figure 12-27: Illustrating the definition of magnetic field intensity at a point in terms of force exerted on a current element; a current-carrying loop of wire is shown (being under the influence of a given set of source currents, not shown in the figure), the force on which can be looked upon as being the vector sum of a large number of elementary forces, each elementary force being exerted on a small portion of the current loop, which we refer to as a length element; one such length element is shown, on which the elementary force $\delta \mathbf{F}$ defines the magnetic field intensity \mathbf{B} at the location of the length element by eq. (12-47).

Once again this equation defining the magnetic field intensity looks more complicated than the corresponding equation defining the electric field intensity in terms of the force on a point charge or an elementary volume element of charge, because of the way the vector cross product appears in it. But that is something one has to live with, though the basic approach underlying the two concepts remains the same. For the sake of completeness, I have to state that the total force on the current loop of which the length element considered in (12-47) is a part, is obtained by summing up all the elementary forces:

$$\mathbf{F} = \sum \delta \mathbf{F}. \quad (12-48)$$

In practice, the summation often takes the form of *integration* along the length of the current loop, which can be worked out exactly in a number of relatively simple situations. For instance, if the magnetic field intensity is *uniform* in a certain region of space

and if a straight segment of a current-carrying wire is placed in that region then the magnitude of the force on a length l of the wire exerted by the field reads

$$F = IBl \sin \theta, \quad (12-49)$$

where θ is the angle between the direction of current through the wire and the direction of the field intensity, the direction of force being perpendicular to both, related to these by the right hand rule of vector cross product (see sec. 2.6), as in fig. 12-28. In other words, the vector expression for the force in this case is

$$\vec{F} = Il\hat{n} \times \mathbf{B}, \quad (12-50)$$

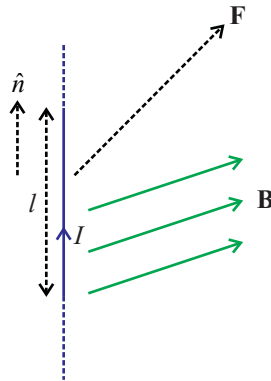


Figure 12-28: Force on a straight portion of a current-carrying wire in a uniform magnetic field of intensity \mathbf{B} ; the force \mathbf{F} is related to \hat{n} and \mathbf{B} by the right-handed rule of vector cross-product, where \hat{n} is the unit vector along the length of the wire in the sense of the current I .

where \hat{n} represents the unit vector along the length of the segment of wire under consideration, in the same sense as the direction of flow of current. Note from equation (12-49) that the magnitude of the force is maximum in a field of given strength when the length of the wire is set at right angles to the direction of the field intensity, in which case one obtains

$$F = IlB. \quad (12-51)$$

This gives a convenient definition of the quantitative measure of the magnetic field intensity: *it is represented in magnitude by the force on a wire of unit length carrying a current of unit magnitude set at right angles to the direction of the field intensity vector.* The unit of magnetic field intensity is termed a *tesla* (T). In accordance with the above definition, it corresponds to the intensity of a magnetic field exerting a force of 1 N on a wire of length 1 m carrying a current of magnitude 1 A, set at right angles to the direction of the field.

12.8.4 The force on a moving charge in a magnetic field

The force on a current-carrying wire in a magnetic field can be looked at from a more fundamental level. Recall that a current is caused by moving charges. Thus the magnetic force on a current can be interpreted, from a fundamental point of view, as the force exerted by a magnetic field on moving charges. Experimental observations have been found to support this interpretation and the formula expressing this force goes by the name of the *Lorentz force* law:

$$\mathbf{F} = q\mathbf{v} \times \mathbf{B}. \quad (12-52a)$$

In this formula \mathbf{F} represents the force on a charge q moving with velocity \mathbf{v} in a magnetic field, the magnetic field intensity at the instantaneous location of the charge being \mathbf{B} . The magnitude of this force is

$$F = qvB \sin \theta, \quad (12-52b)$$

where θ is the angle between the vectors \mathbf{v} and \mathbf{B} . If the two vectors are perpendicular to each other, the formula (12-52b) for the magnitude of the force reads

$$F = qvB. \quad (12-52c)$$

Note that the Lorentz force on a charged particle exerted by a magnetic field *does not perform any work* in a motion undergone by the particle, since the force is perpendicular to the instantaneous velocity, i.e., to the direction of the displacement during a small

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

interval of time at any given instant.

Since the magnetic force on a current-carrying conductor derives from the Lorentz force on moving charges, magnetic forces on currents exerted by steady magnetic fields cannot, in the ultimate analysis, perform work. *Changing magnetic fields*, on the other hand, lead to the generation of induced electromotive forces (see section 13.2) and may be associated with non-zero work performed on currents.

To summarize, a current-carrying wire creates an influence, termed a magnetic field, in the region around itself that causes a force to be exerted on a second current-carrying wire placed in the region; the force is given by formulae (12-43) and (12-45), and the strength of the magnetic field, termed the magnetic field intensity, can be defined in terms of the force by the formulae (12-47), (12-48), which simplifies to (12-50) for a straight segment of wire of length l placed in a uniform magnetic field. The force on a current in a magnetic field derives from the Lorentz force (12-52a) experienced by a moving charge in a magnetic field. The work done by the Lorentz force on a moving charge is zero.

Problem 12-16

A charged particle of charge $q = 1.60 \times 10^{-19}$ C and mass $m = 1.67 \times 10^{-27}$ kg enters into a region of uniform magnetic field $B = 0.5$ T, making an angle $\theta = \frac{\pi}{6}$ with the direction of the magnetic lines of force (i.e., with the direction of the magnetic intensity). If the force experienced by the particle is $F = 2.5 \times 10^{-14}$ N, calculate its kinetic energy in electron volt.

Answer to Problem 12-16

SOLUTION: The Lorentz force on the particle in the magnetic field is given by $F = qvB \sin \theta$, where v stands for the velocity of the particle, and θ for the angle between the direction of the magnetic lines of force and the direction of motion of the particle. This gives $v = \frac{F}{qB \sin \theta}$, and the kinetic energy is $K = \frac{1}{2}mv^2$. With all units in the SI system, this comes out in J. In order to convert this into electron volt (eV), one has to make use of the definition that an eV (refer to section 1.5.3) is the energy change of an electron as it moves through a potential difference of 1 V, i.e., 1 eV equals $q_0 V_0$ J where $q_0 = 1.6 \times 10^{-19}$ C is the magnitude of the charge of an electron and $V_0 = 1$ V. In other

words, in the present instance, $K = \frac{1}{2}m \frac{F^2}{q^2 B^2 \sin^2 \theta} \times \frac{1}{q_0 V_0}$ eV. Using given values, this works out to $K = 2.04$ keV.

The formula $K = \frac{1}{2}mv^2$ is valid only in the framework of *non-relativistic mechanics* we have been considering in this book. If the velocity v happens to be comparable to the speed of light in vacuum, one has to use the formulae of *relativistic mechanics* (refer to chapter 17 for a brief introduction). In the present instance, the non-relativistic formula is adequate (check this out).

12.8.5 Field variables: the question of nomenclature

The naming of the electrical field variables has been discussed in sections 11.10.5.2 and 11.10.5.3. The naming of the magnetic variables is, at times, a source of some confusion. The field variable commonly denoted by the vector \mathbf{B} has been referred to in this book as the *magnetic field intensity* or, in brief, the ‘magnetic intensity’, where the term ‘intensity’ is, once again, in little danger of being confused in practice with the rate of flow of field energy per unit area, the latter being the commonly used meaning of the term ‘intensity’. An alternative designation for the same vector \mathbf{B} that is also in common use is *magnetic flux density*.

Another magnetic vector of considerable relevance is commonly denoted by the symbol \mathbf{H} , and plays a role analogous to the electric displacement vector \mathbf{D} . There is no uniformly accepted term for this vector, though the terms ‘magnetic field strength’ and ‘magnetic field intensity’ are sometimes used (at times, again, the term ‘field strength’ is used for the magnitude of the magnetic field intensity \mathbf{B}). The unit of H is $\text{A}\cdot\text{m}^{-1}$. In this book, however, we will not have much occasion to refer to this magnetic vector.

12.8.6 Field due to a current loop. Principle of superposition.

While the *effect* of a magnetic field is the force on a current-carrying wire placed in it, the field itself is *caused* by one or more current-carrying wires in the first place. The formula for the magnetic field intensity at a point (say, P) due to a wire carrying a current I is once again obtained by summing over contributions coming from all the small length elements (which act as independent sources in generating the field) on the wire. This

fact, that the intensity at a point is the vector sum of contributions made by source elements independently of one another, is the *principle of superposition* in magnetism, similar to the one in electrostatics. The typical elementary contribution to the field intensity coming from a typical element of length, say, δl , is given by the formula

$$\delta \mathbf{B} = \frac{\mu_0 I}{4\pi} \frac{\delta l \hat{n} \times \mathbf{r}}{r^3}, \quad (12-53)$$

where μ_0 is the permeability of free space, \mathbf{r} is the vector distance from the location (say, A) of the length element on the wire to the point P (Fig. 12-29), and \hat{n} is the unit tangent vector at A in the direction of flow of current. Points like A, representing locations of length elements that generate a magnetic field around themselves are referred to as source points while points like P where one calculates the magnetic field intensity are referred to as *field points*. The intensity due to the entire wire at the point P is then given by the vector sum

$$\mathbf{B} = \sum \delta \mathbf{B}, \quad (12-54)$$

and, once again, the summation over a large number of elementary contributions appears as an *integration* over the length of the wire.

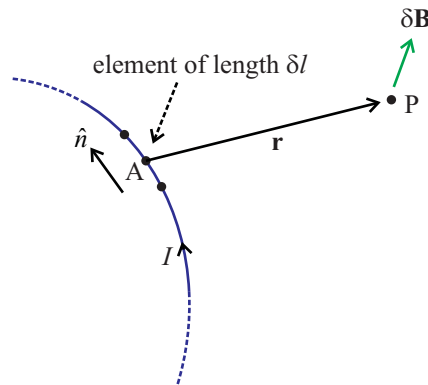


Figure 12-29: Magnetic field due to a current-carrying wire; the wire can be looked upon as being made up of a large number of small length elements, one such element being shown, located at the source point A; the magnetic field intensity at the field point P is the vector sum of a large number of contributions (see fig. 12-30), each coming from a length element like δl ; \hat{n} is the unit vector along the tangent to the contour of the wire at A, in the sense of flow of the current I ; the contribution of this element is $\delta \mathbf{B}$, given by formula (12-53).

Formula (12-53) tells you that the magnetic field intensity at P due to a tiny length element, represented by the vector $\delta\vec{l}$ pointing along the direction of the current and located at any given point A of the wire has magnitude

$$\delta B = \frac{\mu_0 I}{4\pi} \frac{\delta l \sin \theta}{r^2}, \quad (12-55)$$

where θ is the angle between the tangent to the contour of the wire at P pointing along the direction of the current, and the vector \mathbf{r} pointing from A to P. Since the magnetic field intensity is a vector quantity, it will not do to just sum up all these magnitudes to obtain the magnitude of intensity due to the entire wire. Instead, one has to note that the field intensity (12-53) is directed along the unit vector perpendicular to both $\delta\vec{l}$ and \mathbf{r} , which points, in general, along different directions for different locations of the point A on the wire. All the tiny vectors representing the field intensities due to the various length elements on the wire look somewhat like the small arrows in fig. 12-30 which shows schematically the resultant intensity \mathbf{B} at the point P.

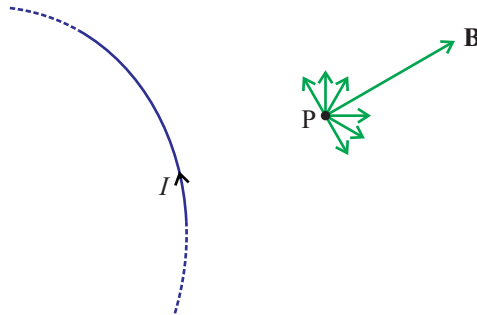


Figure 12-30: Magnetic field intensity due to a current-carrying wire; the intensity is the vector sum of a large number of contributions, each from a small length element of the wire (see fig. 12-29); the small arrows at the field point P represent schematically these contributions, while their vector sum gives the resultant field intensity \mathbf{B} .

An expression for the resultant field intensity due to a closed loop of wire carrying a current I looks like

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} I \oint \frac{d\mathbf{l} \times \mathbf{u}}{u^3}. \quad (12-56)$$

In this expression, dl represents the vector length of an infinitesimal element of closed loop of wire located at, say, the point r' (the source-point), where the direction of the vector is to be taken along the direction of flow of current through the wire, and u stands for the vector $r - r'$, extending from the source point to the field point. In this context, note the change in notation from eq. (12-55) to (12-56). In the former, r was used to denote the vector from the source point to the field point while the same vector is denoted by u in the latter, since r stand nows for the position vector of the field point. The notation in the two expressions, however, is consistent since in eq. (12-53) the origin is taken at the source point (r' in eq. (12-56)).

The integration symbol in the expression (12-56) stands for integration over the closed loop of the current-carrying wire and is defined as the vector sum of a large number of expressions of the form (12-53), each arising from a small element of the wire loop, where the latter is imagined to be partitioned into such small elements.

Imagine now a situation where the magnetic field is created not just by a single current loop, but by a number of such loops. One then obtains the magnetic field intensity at any field point P by making use of the principle of superposition over again: *work out the intensity at P due to each of the current loops independently of the others, and then take the vector sum of the expressions so obtained.*

12.8.7 Magnetic lines of force

The formula (12-53) and the subsequent summation over contributions from all the length elements gives you, in principle, the field intensity at any arbitrarily chosen field point in the magnetic field. The magnetic intensities at all such field points constitute a *vector field* (see, sec. 2.13), and one can think of *magnetic lines of force* to describe geometrically the direction of magnetic field intensity at various points in this vector field.

A magnetic line of force is a line, the tangent to which at every point gives you the direction of the field intensity at that point. In this, a magnetic line of force is the analog of an electrical line of force we encountered in sec. 11.7. You will find this idea

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

of magnetic lines of force illustrated with a number of examples in the following pages where I will also indicate a number of crucial points of difference with electrical lines of force. Lines of force in an electric or a magnetic field are also referred to as *field lines*.

Referring back to (12-53), let me tell you that it is often difficult to work out a formula that can conveniently be employed to obtain the resultant magnetic field intensity at any given point due to a current-carrying wire of arbitrary shape because one has to sum up a large number of contributions by performing an integration over the length of the wire. Such formulae, however, do exist for certain simple situations that I mention below.

Problem 12-17

An electron with kinetic energy $K = 100$ eV enters along the x-axis of a Cartesian co-ordinate system into a region of uniform magnetic field of strength $B = 0.002$ T, with the field lines directed along the y-axis. What is the magnitude and direction of the smallest electric field that will cause the electron to continue to move along the x-axis? .

Answer to Problem 12-17

HINT: Since an electron bears negative charge, the Lorentz force on it due to the magnetic field in the present instance is along the negative z-axis (check this out) and is of magnitude $evB = \sqrt{\left(\frac{2K}{m}\right)}eB$, where v stands for the velocity of the electron, $m(= 9.1 \times 10^{-31})$ kg for its mass, and e for the magnitude of its charge. The condition for the electron to continue moving along the x-axis in the presence of an electric field of strength, say, E , along with the magnetic field, is that the force due to the electric field has to be in the z-x plane, and the z-component of this force has to cancel the Lorentz force.

Recalling that the magnitude of the force due to the electric field is eE , with its direction being opposite to that of E , and assuming that the electric field lines, being parallel to the z-x plane, make an angle θ with the z-axis, the required condition is $-eE \cos \theta = eB\sqrt{\left(\frac{2K}{m}\right)}$, i.e., $E = -\frac{B}{\cos \theta} \sqrt{\left(\frac{2K}{m}\right)}$. The smallest magnitude of the field that meets this condition corresponds to $\cos \theta = -1$ (E is the magnitude of the electric field strength here), and is $E = \sqrt{\left(\frac{2K}{m}\right)}B$. Making use of given values (and the fact that $1 \text{ eV} = 1.6 \times 10^{-19} \text{ J}$), the required electric field intensity is $1.186 \times 10^4 \text{ V}\cdot\text{m}^{-1}$, with

the field lines being along the negative z-direction.

12.8.8 Field intensity due to a straight wire

Consider a *straight* segment of wire carrying a current I as shown in fig. 12-31(A), (B), where the dotted line represents the rest of the closed circuit including a source of EMF. The magnetic field intensity at P due to such a straight portion of a wire is given by the formula

$$\mathbf{B} = \frac{\mu_0 I}{4\pi d} (\cos \theta_1 - \cos \theta_2) \hat{n}, \quad (12-57a)$$

where \hat{n} stands for a unit vector perpendicular to the plane containing the length of the wire and the field point P (the plane of the paper in fig. 12-31) and related to the direction of flow of current by the right hand rule (refer to section 2.8), d is the distance of the point P from the straight portion under consideration, and θ_1 and θ_2 are the angles shown in the figure. In other words, the magnitude of the magnetic field intensity in this case is

$$B = \frac{\mu_0 I}{4\pi d} (\cos \theta_1 - \cos \theta_2), \quad (12-57b)$$

while its direction is perpendicular to the plane of the paper, *into* the plane in fig. 12-31(A), and *coming out* of it in fig. 12-31(B).

Of course, (12-57a) does not represent the magnetic field intensity due to the *whole* of the closed circuit, including the portion represented by the dotted line in fig. 12-31(A), (B). As I have already explained, the principle of superposition tells that the magnetic intensity can be represented as the vector sum of two terms, one coming from the straight part of the circuit and the other from the remaining part, where it may not be possible to work out a simple expression for the latter.

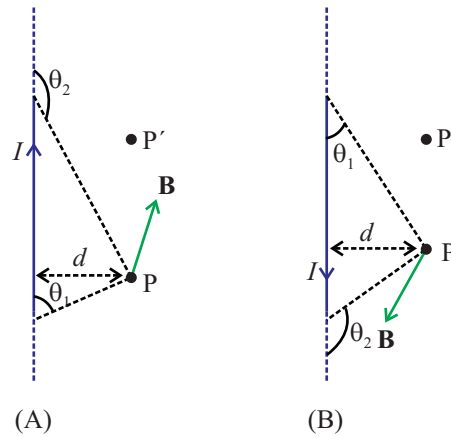


Figure 12-31: Magnetic field intensity due to a straight segment of a wire carrying a current I ; two different directions of current are shown in (A) and (B); the intensity \mathbf{B} is given by formula 12-57a; in (A) it is directed into the plane containing the straight segment under consideration and the field point P , while in (B) it is directed out of the plane; P' represents a second field point located at the same distance from the wire as P ; in general, the magnitude of intensity at P and P' will differ from each other.

12.8.8.1 Infinitely long and straight wire

However, in the special case of a *long* straight wire, it is evidently possible to ensure that this remaining part of the closed circuit is so far removed from the field point (P) under consideration, that the contribution of this part can be ignored. One can imagine an ideal situation where the straight portion of the wire is *infinitely long* and all the small length elements of the second, remaining, part are infinitely away from the field point, in which case the contributions to the field intensity from this remaining part of the circuit actually reduces to zero. In this idealized situation the two ends of the wire are also located at infinitely large distances from the field point and the angles θ_1 , θ_2 attain the limiting values 0 and π respectively (imagine stretching the length of the wire in fig. 12-31(A), (B) so that the end points recede away from P).

One then gets, in this idealized situation of an infinitely long straight wire with the rest of the closed circuit removed to infinitely large distances, the following simple-looking expression for the magnetic intensity at a field point P (refer to sec. 12.8.11.1 for a derivation)

$$\mathbf{B} = \frac{\mu_0 I}{2\pi d} \hat{n}, \quad (12-58)$$

the unit vector \hat{n} being defined as in (12-57a).

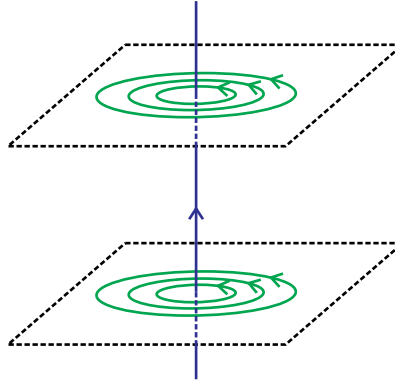


Figure 12-32: Magnetic lines of force due to a straight, infinitely long current-carrying wire; the lines of force are all circular, each contained in a plane perpendicular to the wire; two sets of lines of force are shown in two different planes; the direction of rotation of all the lines of force is related to the direction of the current by the right hand rule.

Notice that, for a given current I , the magnitude of the magnetic field intensity now depends *only* on the distance of the field point from the wire. In the situation depicted in fig. 12-31(A) or (B), for instance, the point P' is located at the same distance from the wire as P , but the magnitudes of the intensity are clearly different for the two points since they correspond to different pairs of values of the angles θ_1 and θ_2 . However, for an infinitely long wire for which the end points have moved out to infinitely large distances up and down the length of the wire respectively, the two angles are $\theta_1 = \pi$ and $\theta_2 = 0$ *no matter where the field point is located*, provided only that it has itself not moved out to an infinitely large distance.

This makes the field strength dependent only on the distance d from the wire : the field strength varies in a direction transverse to the length of the wire, but not in a direction parallel to it.

Taking note of the direction of the magnetic field intensity expressed in terms of the unit vector \hat{n} , one can construct the magnetic *lines of force* (see sec. 12.8.7) due to the infinitely long straight wire which are *circular* in shape, with the planes of the circles perpendicular to the length of the wire. Fig. 12-32 depicts two sets of circular magnetic

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

lines of force in two different planes, both being perpendicular to the length of the wire.

The field strengths at all points on any given circle are the same in magnitude since all such points are at the same distance from the wire, and so are the field strengths at all points on a circle of the same radius located on a different plane. In other words, the field strengths are same at all points on a *cylinder* with the length of the wire as its axis. One describes this by saying that the magnetic field created by an infinitely long straight wire is *cylindrically symmetric*.

The result (12-58) will be derived in sec. 12.8.11.1 by making use of *Ampere's circuital law*.

Problem 12-18

Imagine an infinitely long straight wire to be bent at right angles at a point O so that one half of it extends from the origin O along the positive half of the x-axis of a right handed co-ordinate system while the other half extends along the y-axis. If the wire carries a current $I = 1.5$ A (down to the origin along the x-axis and away from the origin along the y-axis) Calculate the magnetic field intensity at the point (a, a) , where $a = 0.4$ m.

Answer to Problem 12-18

HINT: The magnetic intensities due to both halves of the wire point to the negative direction of the z-axis (draw your own diagram for the problem situation stated) and the magnitudes (B_1, B_2) of the intensities produced by the two halves add up. Moreover, one has, $B_1 = B_2 = \frac{\mu_0 I}{4\pi a} (1 - \cos \frac{3\pi}{4})$.

Using given values, $B = B_1 + B_2 = \frac{2\mu_0 I}{4\pi a} (1 + \cos \frac{\pi}{4}) = 1.28 \times 10^{-6} \text{T}$.

Problem 12-19

Imagine two identical circular loops of wire, each of radius $R = 0.6$ m and each carrying a current $I = 2.0$ A, with their planes parallel to each other and to the z-x plane of a right handed Cartesian co-ordinate system, their centers being at $P(0, -a, 0)$ and $P'(0, a, 0)$ respectively, where $a = 0.001$ m. Estimate the force between the two wire loops. Assume that in each of the loops the direction of flow of current is related to the positive direction of the y-axis according to the right hand rule.

Answer to Problem 12-19

HINT: With reference to any small element of length, say, δl on any of the loops (say, the one with center at the point P; call it L_1), the other loop (L_2) may be taken, in the first approximation, to be an infinitely long wire stretched in both directions since the distance between the loops is small compared to the radius R (reason this out). In other words, the magnetic field intensity at the location of the element δl (indeed, at any point of the loop L_1 under consideration) is $B = \frac{\mu_0 I}{2\pi d}$ where $d = 2a$, the direction of the intensity being perpendicular to the length element, in the plane of the loop (check this out), i.e., parallel to the z-x plane. The force on the element is then $\delta F = IB\delta l$ directed *toward* the loop L_2 , along the positive direction of the y-axis (once again, check this out; draw an appropriate figure). The total force on the loop L_1 is then $F = 2\pi IBR = \frac{\mu_0 I^2 R}{2a}$. Making use of given values, one gets $F = 1.51 \times 10^{-3}$ N (approx). The force on the other loop may be seen to be equal and opposite, as it should be (Newton's third law). In other words, with the currents flowing in the same direction, the force is an attractive one.

12.8.9 Magnetic field intensity due to a circular wire

Consider next a *circular loop of wire* of radius a carrying a current I as shown in fig. 12-33(A), (B). The magnetic field intensity at a point P (not shown in the figure) located on the axis of the circular loop (i.e., a straight line passing through the center of the loop and perpendicular to its plane) at a distance, say, z from the center is given by the formula

$$\mathbf{B} = \frac{\mu_0 I a^2}{2(a^2 + z^2)^{\frac{3}{2}}} \hat{n}, \quad (12-59a)$$

where a stands for the radius of the circular loop and \hat{n} now stands for a unit vector along the axis of the wire related to the direction of flow of current by, once again, the right hand rule, i.e., upward in fig. 12-33(A) and downward in 12-33(B).

The magnitude of the magnetic field intensity is thus given by

$$B = \frac{\mu_0 I a^2}{2(a^2 + z^2)^{\frac{3}{2}}}, \quad (12-59b)$$

and, in particular, the magnitude of intensity at the center is obtained by putting $z = 0$ in this expression:

$$B = \frac{\mu_0 I}{2a}. \quad (12-59c)$$

While this represents the magnetic field intensity in magnitude, the *direction* of the magnetic intensity vector is along the axis of the circular wire, related in the right handed sense to the direction of flow of current.

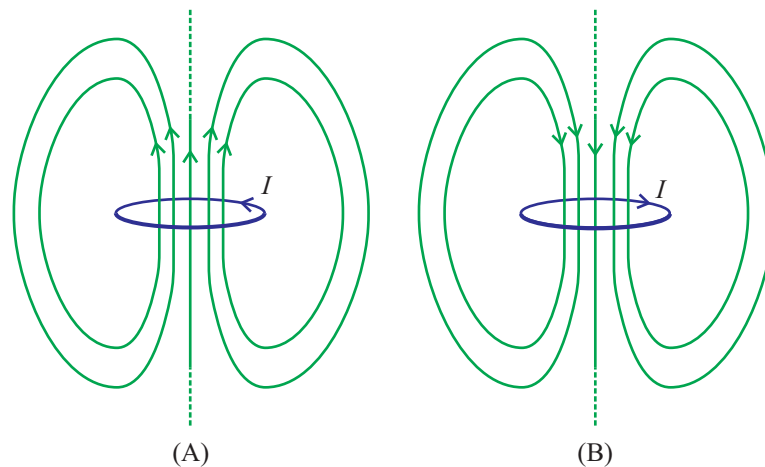


Figure 12-33: Magnetic field due to a circular current-carrying wire; lines of force are shown in (A) and (B) for currents circulating in opposite senses; in each of (A) and (B), one line of force passes through the center of the wire and is a straight line perpendicular to its plane, the directions of the two lines of force being opposite to each other; the field intensity on an axial point is given by the expression (12-59a); for off-axis points, the lines of force are as shown, and are closed lines.

Fig. 12-33 also depicts schematically the lines of force for the circular wire where field lines passing through off-axis points have been shown in addition to the single axial field line passing through the on-axis points. Though there does not exist a simple expression for the intensity at off-axis points, lines of force can still be drawn, say, with the help of an appropriate computer program. Notice that the axial line of force is a straight line piercing the plane of the circular wire, running from ‘minus infinity’ to ‘plus infinity’, while each of the other lines of force is a closed curve that pierces the plane, curves away from the axis, turns back, and then closes upon itself. The single

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

on-axis line of force does not appear to close on itself, but this is an exception rather than the rule, and one can imagine that it ‘closes at infinity’.

The expressions (12-59b) and (12-59c) apply to the magnetic field intensity due to a single turn of a circular wire, though one can easily generalize these for a planar circular coil of N turns where, for simplicity, all the N turns are assumed to be coincident circles of the same radius. In reality, of course, no single turn of the coil forms a complete circle and the successive turns deviate from a single circular contour. However, for a tightly wound coil, the contribution to the intensity by each turn of the coil is, to all intents and purposes, the same as that due to a circular wire of single turn. Thus, the intensity due to the coil as a whole is simply N times the intensity due to a single turn, and the field strength at the center is then

$$B = \frac{\mu_0 N I}{2a}. \quad (12-60)$$

Problem 12-20

A pair of circular coils, of radii $a_1 = 0.2\text{m}$ and $a_2 = 0.15\text{m}$, is placed as in fig. 12-34, with their centers on the x- and y-axes of a right handed Cartesian co-ordinate system, at points $(u_1 = 0.2, 0, 0)\text{m}$ and $(0, u_2 = 0.3, 0)\text{m}$, the planes of the coils being perpendicular to the two axes. The coils carry currents $I_1 = 1.5\text{A}$, $I_2 = 2\text{A}$, where the directions of the currents are related to the positive directions of the x- and y-axes by the right hand rule. If the numbers of turns of the coils be, respectively, $N_1 = 50$, $N_2 = 100$, obtain the magnetic field strength at the origin O.

Answer to Problem 12-20

HINT: The magnetic field intensity due to a coil of a single turn being given by expression (12-59b) (with notation explained along with the formula), the field intensity due to N turns, all assumed to be of the same radius, is obtained by multiplying this expression with N . Thus, from the data provided, the magnetic field intensities due to the two circular coils are of magnitudes $B_i = \frac{\mu_0 N_i I_i a_i^2}{2(a_i^2 + u_i^2)^{\frac{3}{2}}}$ ($i = 1, 2$), these being directed along \hat{i}, \hat{j} respectively, i.e., along the unit vectors in the positive directions of the two axes. Substituting given values, the resultant field intensity is seen to be, by the principle of superposition, $\mathbf{B} = B_1 \hat{i} + B_2 \hat{j} = 10^{-5} \times (8.32\hat{i} + 7.47\hat{j})\text{T}$. In other words,

the magnitude of the field intensity is $B = 11.18 \times 10^{-5} \text{T}$, and the direction of the field makes an angle $\theta = \arctan \frac{7.47}{8.32} = 0.73$ radian with the positive direction of the x-axis, pointing into the first quadrant of the x-y plane, as indicated in fig. 12-34.

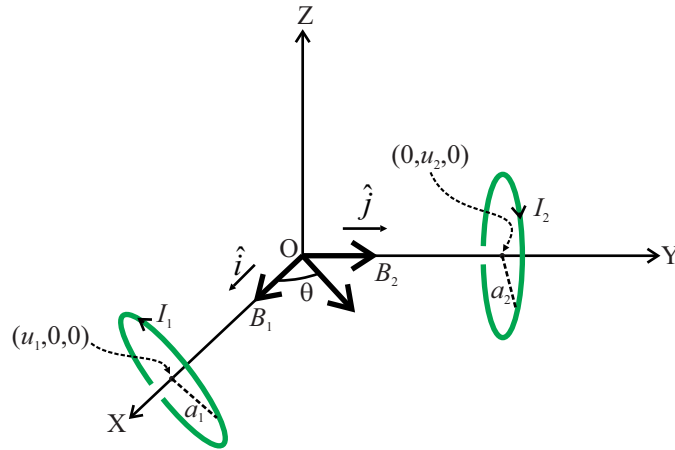


Figure 12-34: Two circular coils of wire with numbers of turns N_1 , N_2 , and of radii a_1 , a_2 , placed with their centers on the x- and y-axes of a right handed Cartesian co-ordinate system, at points $(u_1, 0, 0)$ and $(0, u_2, 0)$, the planes of the coils being perpendicular to the two axes; the coils carry currents I_1, I_2 , where the directions of the currents are related to the positive directions of the x- and y-axes by the right hand rule; by the principle of superposition, the resultant field intensity at the origin is $\mathbf{B} = B_1 \hat{i} + B_2 \hat{j}$, where B_i ($i = 1, 2$) are obtained as in problem 12-20; the magnitude of the field intensity is $B = (B_1^2 + B_2^2)^{\frac{1}{2}}$, while \mathbf{B} makes an angle θ with the positive direction of the x-axis.

12.8.9.1 Magnetic field of a solenoid

Finally, consider a tightly wound *solenoid*, i.e., a coil made up of circular turns, all of the same radius, where each turn is displaced slightly from the adjacent ones in a direction parallel to the common axis of the circles like, for instance a wire wound on a wooden cylinder, with little gap in between successive turns of the wire (fig. 12-35). The axis of the cylindrical frame is termed the axis of the solenoid. Let P be a point on the axis. Choosing two points A and B on the end faces of the cylindrical frame as in the figure, let θ_1 and θ_2 be the angles made by the lines PA and PB with the axis, as shown. Let N denote the number of turns in the solenoid and l its length, so that the number of turns per unit length of the solenoid is $\frac{N}{l}$. Then the magnetic intensity at P due to a current I

in the solenoid is given by the expression

$$\mathbf{B} = \frac{\mu_0 N I}{2l} (\cos \theta_2 - \cos \theta_1) \hat{n}, \quad (12-61a)$$

where \hat{n} stands for the unit vector parallel to the axis of the solenoid related to the direction of flow of current by the right hand rule. The magnitude of the field intensity on an axial point of a solenoid is thus

$$B = \mu_0 \left(\frac{N}{2l} \right) I (\cos \theta_2 - \cos \theta_1). \quad (12-61b)$$

It is of interest to work out the magnetic field strength at an interior point on the axis of a *long* solenoid since such long solenoids are often employed to generate *uniform magnetic fields*. The geometry of this situation is to be inferred by extrapolation from fig. 12-35 where the point P is located on the axis in between the two end faces of the cylindrical frame and, the solenoid being a long one, the end faces are at a *large* distance on either side of P. In other words, unless the point P is located close to one of the end faces, the angles θ_1 and θ_2 are close to π and 0 respectively.

For many situations of interest one can even go over to the limit of an *infinitely long* solenoid for which θ_1 and θ_2 can be set at π and 0 respectively for *all* interior points on the axis sufficiently removed from either end faces. Though, in this limit, N and l both have infinitely large values, the ratio $\frac{N}{l}$ is a well defined, finite quantity, being the number of turns *per unit length* of the solenoid, which we now denote by, say, n_0 . One then ends up with the following simple expression for the magnetic field intensity at an interior point on the axis of a tightly wound, infinitely long solenoid:

$$\mathbf{B} = \mu_0 n_0 I \hat{n}, \quad (12-61c)$$

where \hat{n} is once again a unit vector directed along the axis of the solenoid, related to the direction of flow of current in the windings in the right hand sense. A derivation of this result using *Ampere's circuital law* will be presented in sec. 12.8.11.

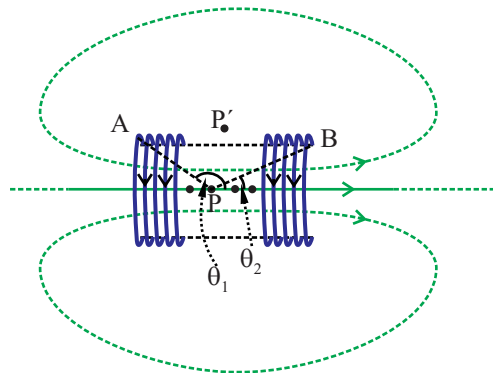


Figure 12-35: Magnetic field due to a solenoid; only a few turns of the solenoid near the two ends are shown (other turns are represented by dots); the field strength at an axial point like P is given by expression (12-61a); a few lines of force are also shown; one of these extends along the axis to infinite distances on both sides; others bend away from the axis and close on themselves; for a point like P' outside the solenoid and lying close to the cylindrical surface on which the solenoid is wound, the field strength is low, becoming vanishingly small for an infinitely long solenoid; for a long solenoid, the off-axis lines of force run almost parallel to the axis up to large distances before turning back and closing upon themselves; consequently, no line of force passes through a point like P'.

Fig. 12-35 shows schematically a number of lines of force describing the magnetic field set up by a solenoid. One of these lines of force lies along the axis of the solenoid while the others bend away from the axis, closing on themselves. For a long solenoid the lines of force are almost parallel to the axis in the interior and keep on running parallel to the axis even as they emerge on either side from the interior, bending away from the axis only at large distances.

There are a few interesting aspects to the magnetic field generated by a long solenoid. For one thing, the field strength is the *same* at *all* interior points on the axis sufficiently removed from the two end faces, being of value $\mu_0 n_0 I$. What is more, it has the *same value at off-axis interior points as well* where, once again, the points are assumed to be sufficiently removed from the two end faces. In other words the long solenoid generates a *uniform magnetic field* throughout its interior region, barring regions close to the end faces.

Finally, the field strength at any point like P' just outside the solenoid works out to *zero*. The magnetic lines of force representing the field generated by the long solenoid resemble the ones shown in fig. 12-35, but the resemblance is not apparent in the

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

interior region of the solenoid, as also in the exterior region close to it. These lines of force are parallel to the axis throughout the volume inside the cylindrical frame (in most cases an insulating material body of cylindrical shape, but may as well be an imaginary cylindrical surface, the turns of the wire being then held together by some glue-like substance) barring the regions close to the two ends, corresponding to the fact that the magnetic field intensity is parallel to the axis at all points in this interior region.

In the exterior region, the magnetic lines run almost parallel to the axis towards faraway points on the two sides, bending only slightly away from one another. The bending increases at large distances from the two ends of the solenoid, where the lines turn backwards so as to form closed paths as shown schematically in figure 12-35 for some of the lines of force. For other lines, closer to the axis, the bending occurs at larger distances. It follows that there are no lines of force running backwards (representing an intensity in the opposite direction as compared to the intensity at interior points) close enough to the cylindrical surface, consistent with the fact that the field strength is zero at such points.

Incidentally, the basic rules for calculating magnetic field strengths for given current loops can be invoked to work out the magnetic field intensity at any point *on* the end face of a closely wound long solenoid, which is given by the expression

$$\mathbf{B} = \frac{1}{2}\mu_0 n_0 I \hat{n}, \quad (12-62)$$

which is to be compared with the expression (12-61c). In other words, the field strength is given by (12-61c) for a point in the interior the solenoid away from either end face, by the expression (12-62) for a point on either end face, and is zero at a point external to the solenoid, once again away from the end faces. These results hold in a limiting sense for a long and closely wound solenoid. In reality, however, there arises small but finite deviations from these results.

While the expressions for magnetic field intensity in these few (and some other) special cases are simple ones, simple-looking formulae cannot be written down for other situa-

tions like, for instance, for the field intensity due to a circular wire at an *off-axis* point, or the intensity at an off-axis point for a solenoid of finite length. However, recall that in all such situations, the intensity is given by an integral expression like (12-54), i.e., a sum over a large number of infinitesimal contributions of the form (12-53) coming from infinitesimal length elements of the current-carrying circuit generating the field, and so there is no problem *in principle* in defining and calculating the field intensity - for instance, a computer program can be used to make such a calculation numerically. What one really needs is the expression (12-53), together with the superposition principle, and the rest is nothing but the summing up of the elementary field contributions.

It does not even matter if the field is generated by one single current-carrying circuit or more than one circuits because, as I have mentioned earlier, knowing the field generated by each circuit separately, one just needs to apply the superposition principle once again to work out the resultant field generated by all the circuits taken together.

12.8.10 Ampere's circuital law

Figure 12-36 depicts parts of two closed loops of wire carrying currents I_1 and I_2 (the rest of the loops are not shown in the figure), and a closed path C encircling the two currents. There may be other current loops (not shown in the figure) contributing, along with the loops carrying the currents I_1 and I_2 , to the magnetic field intensity at any given point, but we assume, for the sake of illustration, that these other currents do not cross the chosen closed path C. The field intensity at any given point \mathbf{r} can be worked out by invoking the basic formula (12-56) along with the principle of superposition. Considering any chosen sense of traversal of the path C (say, the one shown by the double-headed arrow in the figure), one can define the line integral of the magnetic field intensity \mathbf{B} along the path, by imagining the path to be partitioned into a large number of infinitesimally small segments, and then summing up contributions of the form $\mathbf{B} \cdot \delta \mathbf{l}$ from these segments, where \mathbf{B} denotes the field intensity at the location of a typical segment and $\delta \mathbf{l}$ stands for the vector length of the segment, taken along the sense of traversal of the path C.

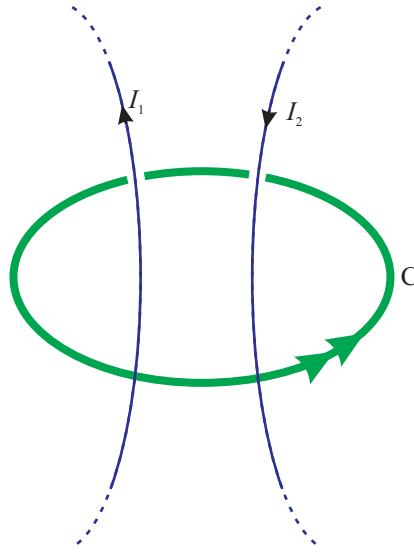


Figure 12-36: Illustrating Ampere's circuital law; C is a closed path in a magnetic field, where it may be the contour of a conducting wire or even an imagined path; among the current loops generating the magnetic field, the contour of two are as shown (only parts of the two closed loops are drawn in the figure), piercing the closed path C ; the other current loops (not shown) do not cross the closed path C ; the sense of traversal of the current I_2 is shown to be opposite to that of I_1 where, more generally, there may be any number of currents crossing the path C ; a sense of traversal of C is chosen, say, along the double-headed arrow; Ampere's circuital law then relates the line integral of the magnetic field intensity along the closed path C (with the integral taken along the chosen sense of traversal) with the algebraic sum of the currents crossing the closed path C .

Ampere's circuital law states that the line integral of the magnetic field intensity along any such path C is μ_0 times the algebraic sum of the currents encircled by C , where each of the currents is to be taken with its appropriate sign, which is to be positive if the chosen sense of traversal of C is related to direction of the current by a right hand rule:

$$\oint_C \mathbf{B} \cdot d\mathbf{l} = \mu_0 \sum I. \quad (12-63)$$

What is important to note is that *the contribution to the above line integral of the current loops that do not cross the path C is zero*, which means that the sum on the right hand side of eq. (12-63) includes *only* the currents crossing the closed path under consideration (I_1 and I_2 in the present instance), though the field intensity \mathbf{B} occurring in the integral on the left hand side is obtained by superposing the field intensities due to *all* the current loops, regardless of whether or not they cross the contour C . Thus, applied

to the situation depicted in the figure 12-36, the law gives

$$\oint_C \mathbf{B} \cdot d\mathbf{l} = \mu_0(I_1 - I_2), \quad (12-64)$$

where I_2 occurs with a negative sign in the right hand side since the direction of I_2 is related to the chosen sense of traversal of the path C in a left handed sense.

Ampere's circuital law is a principle in magnetism having a status analogous to the Gauss principle in electrostatics.

In writing the formula (12-63) I have implicitly assumed that the closed loop under consideration is located in vacuum. More generally, if the loop is located in a material medium, the circuital law takes the form

$$\oint_C \mathbf{B} \cdot d\mathbf{l} = \mu_r \mu_0 \sum I, \quad (12-65)$$

where μ_r is a constant referred to as the *relative permeability* of the medium. For a large number of media, however, μ_r happens to be close to unity, as a result of which one may continue using the formula (12-63) without the possibility of appreciable error. In the case of the so-called *magnetic* materials (more precisely, the *ferromagnetic* ones; see sec. 12.9 for a brief introduction to the magnetic properties of materials), on the other hand, μ_r may differ significantly from unity and may, moreover, depend on the magnetic field strength in the medium.

12.8.11 Applications of the circuital law

As with the Gauss' principle in electrostatics, Ampere's circuital law can be made use of in conveniently working out the magnetic field intensity in a number of symmetric situations.

12.8.11.1 Ampere's law: field due to a long straight wire

Consider, for instance, the magnetic field due to an infinitely long and straight current-carrying wire (sec. 12.8.8.1), with the field point P chosen at a distance d from the wire.

In this case, the path C is to be chosen as in fig. 12-37: C is a circle passing through P , imagined to be drawn in a plane perpendicular to the length of the wire, with the center of the circle located on the wire. The symmetry of the situation implies that B has a constant magnitude at all points on the circle, while being directed tangentially to it (reason out this statement). We choose the sense of traversal of the circle (double-headed arrow) as the one related to the direction of the current by the right hand rule, and assume that the magnetic field intensity at any point on the circle along the tangential direction consistent with the chosen sense of traversal is B . The equation (12-63) then reduces to the form

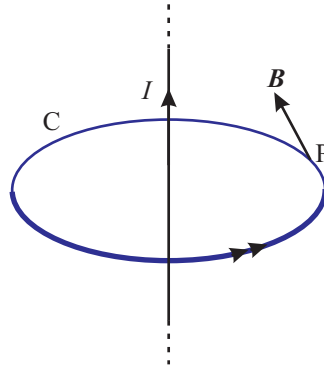


Figure 12-37: Illustrating Ampere's circuital law for a long straight wire; P is a point in the field created by the current in the wire: a closed path C in the form of a circle passing through P and having the straight wire as its axis, is chosen, along with a sense of traversal shown by the double-headed arrow; the circuital law then gives the magnetic field intensity at any point on C in the form of the expression (12-66).

$$2\pi B d = \mu_0 I, \text{ i.e., } B = \frac{\mu_0 I}{2\pi d}, \quad (12-66)$$

entirely in accord with the result (12-58).

12.8.11.2 Ampere's law: the tightly wound long solenoid

As another application, consider a tightly wound infinitely long solenoid. In section 12.8.9.1 I pointed out that the magnetic field intensity is uniform everywhere in the interior of such a solenoid and zero at any exterior point, since the field lines get bent and turn back (as in fig. 12-35) at infinitely large distances away from the solenoid. Assuming that the magnitude of the magnetic field intensity along the axis of the solenoid (in a direction related to the sense of the solenoid current by the right hand rule) is B , the expression (12-61c) for B can be worked out by referring to fig. 12-38 given below.

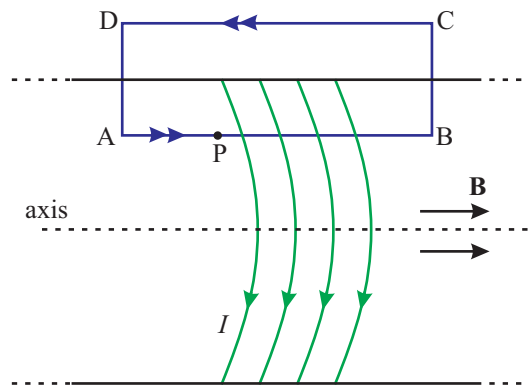


Figure 12-38: Illustrating Ampere's circuital law for a tightly wound long solenoid.

The figure shows the cylindrical surface supporting the winding of the solenoid (this may be an imaginary surface, though more often it is the surface of a cylinder made of a non-conducting material; I assume the solenoid to be an *air-cored* one, while solenoids with a magnetic material introduced inside the cylindrical surface are also common). Only a few turns of the solenoid winding are shown schematically.

Considering any given field point P in the interior, the rectangle $ABCD$ depicts a closed path with two of its sides parallel to the axis of the solenoid, of which one (AB) lies in the interior, passing through P , and the other (CD) in the exterior region. The two remaining sides (BC and DA) of the path are perpendicular to the axis.

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

On applying Ampere's circuital law to this path, with the direction of traversal related by the right hand rule to the direction of the current through the winding (see figure), the left hand side of eq. (12-63) can be broken up into four parts corresponding to the four arms of the rectangular path, of which only the arm AB contributes to the integral: the contribution of the arm CD is zero because the magnetic field intensity is zero at any exterior point, while those of BC and AD are both zero since the magnetic field intensity is perpendicular to any length element on either of these arms. As regards the sum on the right hand side of (12-63), each turn of the solenoid enclosed by the path ABCD makes a contribution I to the sum, where I stands for the solenoid current. If n_0 is the number of turns per unit length of the solenoid, and l be the length of the arm AB, then the number of turns enclosed by the path ABCD is $n_0 l$, and thus one arrives at

$$Bl = \mu_0 n_0 l I, \text{ i.e., } B = \mu_0 n_0 I, \quad (12-67)$$

entirely in accord with the result (12-61c), where the latter gives the vector expression for the magnetic field intensity.

Problem 12-21

Consider a long cylindrical conductor of radius $R = 0.002$ m made of a non-magnetic material ($\mu_r \approx 1$) carrying a current $I = 4.5$ A flowing parallel to the axis of the cylinder, where the current is uniformly distributed over its cross section. Calculate the magnetic field intensity \mathbf{B} at a point within the material of the cylinder at a distance $r = 0.001$ m from the axis.

Answer to Problem 12-21

HINT: Imagine a closed circular loop (commonly referred to as an 'Ampere loop') through the point under consideration, the plane of the loop being perpendicular to the axis of the cylinder, with the centre lying on the axis. Due to the symmetry of the problem, the magnetic intensity has to be the same in magnitude at all points on the loop, its direction being tangential to the loop, related to the direction of flow of current by the right hand rule. Applying Ampere's circuital law to the

loop, one gets $2\pi rB = \mu_0 I'$, where I' , the current passing through the loop is given by $I' = \pi r^2 j$, j being the magnitude of the current density. Since the latter is uniform over the cross section of the cylindrical conductor, one has $j = \frac{I}{\pi R^2}$. Combining all these formulae, one gets $B = \frac{\mu_0 I r}{2\pi R^2}$. Making use of given values, $B = 2.25 \times 10^{-4}$ T.

12.8.11.3 Infinitely long cylindrical current distribution

The situation described in problem 12-21 can be generalized to the case of an infinitely long cylindrical current distribution, where the current density is every where directed along the z-axis of a co-ordinate system, is independent of the z-coordinate (i.e., is constant everywhere on any line parallel to the z-axis) and of the azimuthal angle about the z-axis, and depends only on the distance ρ from the z-axis, being of magnitude $J(\rho)$ at a distance ρ . In this case, the magnetic field lines are all circular ones, in planes perpendicular to the z-axis, the latter passing through the centers of all the circular field lines, and the magnetic field intensity vector at any point is directed along the tangent to the field line passing through that point, being related to the positive direction of the z-axis in the right handed sense. Let the magnitude of the field intensity, which depends only on ρ be denoted by $B(\rho)$.

An application of the circuital law, in a manner analogous to that in problem 12-21 then gives the following result:

$$B(\rho) = \frac{\mu_0}{\rho} \int_0^\rho J(\rho') \rho' d\rho', \quad (12-68)$$

where ρ' is an integration variable ranging from zero to ρ (check this expression out). In particular, for a uniform current distribution confined within a cylindrical region of cross-section a , for which

$$\begin{aligned} J(\rho) &= \frac{I}{\pi a^2} \quad (0 \leq \rho \leq a), \\ &= 0 \quad (\rho > a) \end{aligned} \quad (12-69)$$

one obtains, by an application of the circuital law (choosing an Ampere loop in the form

of a circle of radius ρ with its plane perpendicular to the z -axis)

$$\begin{aligned} B(\rho) &= \frac{\mu_0 I \rho}{2\pi a^2} \quad (0 \leq \rho \leq a), \\ &= \frac{\mu_0 I}{2\pi a} \quad (\rho > a). \end{aligned} \quad (12-70)$$

Here I is the total current within the cylindrical region of radius a , which is assumed to be located in free space. Note that the expression for B in the region $\rho > a$ is consistent with the formula (12-66).

12.8.12 The magnetic dipole

Consider a tiny current loop which is so small in size that it can be taken to be a planar one (i.e., the contour of the loop is contained in a plane). Let the area enclosed by the loop be A , the current through the loop be I , and the unit normal to the plane of the loop, related to the direction of the current by the right hand rule, be \hat{n} . Let, moreover, the position vector of the loop, which is so small that it can be looked upon as a point source, be \mathbf{r}' .

One can work out the magnetic field intensity at any field point \mathbf{r} away from the location of the current loop by applying the basic formula (12-56), and then consider a limiting form of this formula with $A \rightarrow 0$, i.e., with the loop shrinking to a point, and, at the same time, with $I \rightarrow \infty$ in such a manner that the product AI has a definite limit, say, m .

The interesting result that comes out of such an exercise is that, the limiting expression for the magnetic field intensity \mathbf{B} at the field point turns out to be *independent of the shape of the loop* one starts with and, for given source- and field points, is determined solely by the limiting value of the product AI and the unit vector \hat{n} . This expression looks like

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \frac{3(\mathbf{m} \cdot \hat{u})\hat{u} - \mathbf{m}}{u^3}, \quad (12-71a)$$

where

$$\mathbf{m} = m\hat{n}, \quad (12-71b)$$

and \hat{u} and u are defined with reference to the vector \mathbf{u} extending from the source point to the field point -

$$\mathbf{u} = \mathbf{r} - \mathbf{r}', \quad u = |\mathbf{u}|, \quad \hat{u} = \frac{\mathbf{u}}{u}. \quad (12-71c)$$

It may be mentioned that the expression (12-71a) holds everywhere in space *except* at $u = 0$, i.e., at the location of the dipole. This means that, at the location of the dipole, the field intensity diverges in a manner different from that implied by the above expression. Referring to the above limiting situation, the current loop, acting as a source of the magnetic field with field intensity \mathbf{B} , is termed a *magnetic dipole* of *dipole moment* (or *magnetic moment*) \mathbf{m} .

Note that the expression (12-71a) for \mathbf{B} is analogous to the expression (11-35) for the electric field intensity due to an electric dipole of dipole moment \mathbf{p} where, in eq. (11-35), the vector from the location of the dipole to the field point is denoted by \mathbf{r} instead of \mathbf{u} . Indeed, magnetic dipoles and dipole moments play an analogous role in the description of magnetic fields and their sources as do electric dipoles and dipole moments in the description of electric fields and their sources.

However, while the expressions for the electric and magnetic field intensities due to an electric and a magnetic dipole are analogous, there is, at the same time, a subtle distinction between the two when one looks at the field intensity *at the location of the dipole* in each case. The manners in which the field intensities *diverge* (i.e., becomes infinitely large) at the points of location of the dipoles differ from each other (I do not enter here into further details). This difference between the two situations relates ultimately to the basic difference between electric and magnetic fields (see sections 12.8.12.1 and 12.8.13).

12.8.12.1 Electric and magnetic dipole moments

In sec. 11.6 we distinguished between a real and an ideal electrical dipole. While an ideal electrical dipole involves the limit $d \rightarrow 0$, $q \rightarrow \infty$, a real dipole is made up of a pair of equal and opposite charges placed a small but finite distance apart, and the expressions for the electric potential and field intensity due to this real dipole differ from the corresponding expressions for an ideal dipole by small correction terms which can be ignored in the first approximation.

In an exactly analogous manner, a current loop of small but finite area is commonly referred to as a (real) magnetic dipole, and the magnetic field intensity due to such a real dipole can be approximated by the expression of field intensity due to an ideal dipole. The dipole moment of the real magnetic dipole is defined as

$$\mathbf{m} = AI\hat{n}, \quad (12-72)$$

where A and \hat{n} are now defined in an analogous but approximate manner since the current loop can no longer be assumed to be strictly a planar one. The dipole is said to have its *axis* along the vector \hat{n} .

We saw in chapter 11 that charges can be considered to be the elementary sources of an electric field since, knowing the charge distribution, one can obtain the electrical potential and field intensity at any given field point. What distinguishes a magnetic field from an electrical one is that, according to our present state of knowledge, there exists no such thing as a ‘magnetic charge’. Instead, magnetic dipoles can be looked upon as the elementary sources producing a magnetic field. This provides us with an alternative description of the sources producing a magnetic field: *a magnetic field produced by a current distribution can also be looked upon as being produced by an equivalent distribution of magnetic dipoles.*

All static electric fields in nature can ultimately be described as being produced by elementary charges at the microscopic, or atomic level. In an exactly analogous manner, all static magnetic fields can be described as being produced by elementary magnetic

dipoles of a microscopic nature, where these dipoles are made up of tiny current loops at the atomic and molecular levels. In a following section (section 12.9) we will see that the macroscopic magnetic properties of materials can be explained on the basis of these microscopic magnetic dipoles.

At the microscopic level, a tiny current loop corresponds to the *orbital* motion of charges in the nuclei, atoms, or molecules in a material. Thus, it would seem as if magnetic dipoles are associated with orbital motions of charges of a microscopic nature. However, sensitive measurements reveal that, at the microscopic level, magnetic dipole moments are not accounted for solely by orbital motions of charges since there remain *intrinsic* magnetic moments which are not explained completely in terms of orbital motions. Thus, the ultimate origin of magnetic moments seems to be a deep question, linked up with the question of elementary constituents of matter and their interactions (see section 18.8.9 for an introduction to a few relevant concepts).

Gyromagnetic ratio.

On working out the angular momentum \mathbf{L} and the magnetic dipole moment \mathbf{m} of a particle of mass M and charge q moving in a circular orbit with angular velocity ω , it is found that

$$\mathbf{m} = \frac{q}{2M} \mathbf{L}. \quad (12-73)$$

The quantity $\frac{q}{2M}$ expressing the ratio of the magnetic moment and the angular momentum is termed the *gyromagnetic ratio* of the particle (at times, the gyromagnetic ratio is defined as $|\frac{q}{2M}|$). For a negatively charged particle such as an electron, the direction of the magnetic moment is opposite to that of the angular momentum.

Magnetic dipole moment of a current distribution.

As we saw in sec. 11.6, not only an ideal electrical dipole, but *any* charge distribution is characterized by a dipole moment, and this dipole moment provides us with a convenient approximation for expressing the electric potential and field due to the charge distribution at a *large* distance from the charge distribution. More precisely, if the total

charge of the distribution (the ‘monopole’ moment) vanishes, then the dipole moment accounts for a good approximation to the electric potential and field intensity.

In an exactly analogous manner, a magnetic dipole moment can be associated with not only an ideal magnetic dipole, but with *any* localized current distribution producing a magnetic field. And it is this dipole moment that provides us with a good approximation to the field intensity at *large* distances from the current distribution (described by a current density $\mathbf{j}(\mathbf{r})$ depending on the position vector \mathbf{r}). What distinguishes the magnetic field from the electric one is that, in the case of a current distribution, no ‘monopole’ term occurs in the expression of the magnetic field intensity, expressed in the form of a series with terms of successively smaller orders of magnitude corresponding to increasing inverse powers of the distance from the current distribution. In other words, the dipole moment *always* gives a good approximation to the magnetic field intensity at large distances, being the first term in this series expansion, in contrast to the corresponding series expansion for the electric field intensity where the dipole moment gives the *second* term, coming after the monopole term.

12.8.12.2 Current loop: a surface distribution of dipoles

Fig. 12-39(A) shows a closed loop C of current in a wire (the source of EMF is not shown in the figure) producing a magnetic field, where the current (I) in the loop flows in the direction shown by the arrow.

One can now imagine a surface S whose boundary coincides with the current loop, where any one of the many possible surfaces satisfying this requirement can be chosen. This surface can then be imagined to be made up of a large number of tiny area elements, where the boundary of each of these elements can be imagined to constitute a closed current loop, carrying a current I in the same sense as the current in the loop C we started with.

The common boundary between any two contiguous elements then carries equal currents in opposite directions which cancel each other, as in the segments A and B in fig. 12-39(B). The only currents that do not get canceled are the ones in the outer bound-

aries of the area elements at the edge of the loop C, such as the current in the segment D in fig. 12-39(B).

Thus, the combined effect of all these tiny current loops has to be the same as the effect of the current I in the loop C since the latter is made up precisely of the segments like D. Considering any one of the small current loops of area, say δA , it corresponds to a dipole of moment $I\delta A$ pointing along the normal to the surface S at the point of location of the loop, the sense of the normal being related to the direction of flow of current in the loop C by the right hand rule.

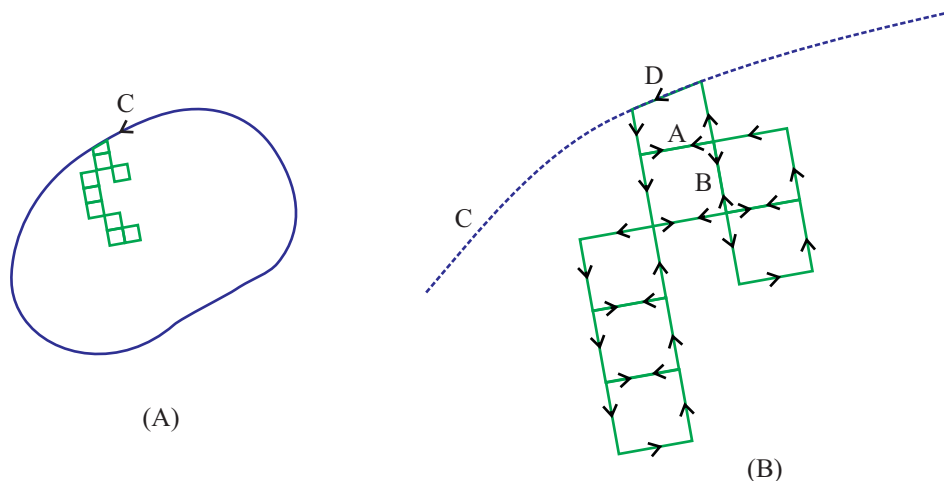


Figure 12-39: (A) Current loop C imagined to be divided into area elements, each carrying a current I ; (B) a few of the elements shown separately in magnification, including one at the edge of C; the currents in each arm of a typical area element all cancel out, leaving only the currents in the outer edges of these boundary elements (such as the edge D), where these residual currents make up the current through C shown in (A).

In other words, the magnetic field set up by a current loop C carrying a current I is equivalent to that of a distribution of dipoles on a surface S spanning the loop (i.e., a surface whose boundary coincides with C; there can be different possible choices for S, any one of which can be adopted), the dipole moment per unit area of the surface being I , with the direction of the dipole axis as described above.

1. In giving you the theory underlying the production and description of stationary

magnetic fields, I started from *currents* as the sources of magnetic fields. We now see that a magnetic field produced by currents can equally well be described in terms of distributions of magnetic dipoles. A current in a conductor producing a magnetic field is caused by the drift motion of *free* electrons in the conductor. There exist, on the other hand, *bound* electrons in all materials that constitute tiny current loops at the microscopic level, which can also produce magnetic fields. These current loops are also equivalent to little dipoles, in terms of which one can explain the magnetic properties of materials. What is more, there possibly exist magnetic moments at the microscopic level that *cannot* be explained in terms of current loops. In this sense, magnetic dipoles may be said to constitute the ultimate sources of magnetic fields where a description in terms of currents as the sources of magnetic fields may prove to be inadequate.

2. When you look at the magnetic field produced by a current in a closed loop of wire, you describe it in terms of the current caused by the drift motion of the electrons in the wire. As we have seen, the same magnetic field can be described by a distribution of dipoles corresponding to a collection of imagined tiny current loops. When, on the other hand, you look at the magnetic field due to a permanent magnet, you describe the field in terms of a distribution of dipoles once again, where these dipoles correspond to tiny current loops actually circulating at the atomic level. In either case, magnetic dipoles can be looked upon as the ultimate sources producing magnetic fields. The reason the dipole picture is superior over the picture involving currents is two-fold: (a) the dipole description is much more convenient compared to the current description for situations such as the one involving a permanent magnet, and (b) there may possibly exist magnetic dipoles at the level of elementary constituents of matter that may not be amenable to a description in terms of current loops caused by the orbital motions of charges. This is indicated by *anomalous values* of the gyromagnetic ratio in the case of a number of elementary constituents as compared to the value $\frac{q}{2m}$ (see eq. (12-73)), and relates to magnetic moments of an *intrinsic* nature. For instance, the neutron is an uncharged particle while, at the same time, possessing a magnetic moment of its own.

12.8.12.3 Torque and force on a magnetic dipole

Consider a small current-carrying loop, constituting a magnetic dipole of dipole moment, say, \mathbf{m} , placed in a magnetic field \mathbf{B} , which may possibly be a non-uniform one. Imagining the contour of the loop to be divided into a large number of length elements, one can work out the magnetic force on all these length elements, and then calculate their resultant. In the limit of the current loop being reduced to an ideal dipole, the resultant is equivalent to a single force, say, \mathbf{F} acting on the dipole (i.e., with its line of action passing through the point of location of the dipole) together with a couple of moment, say, \mathbf{M} , referred to as the *torque* on the dipole, where

$$\mathbf{F} = \hat{i} \frac{\partial(\mathbf{m} \cdot \mathbf{B})}{\partial x} + \hat{j} \frac{\partial(\mathbf{m} \cdot \mathbf{B})}{\partial y} + \hat{k} \frac{\partial(\mathbf{m} \cdot \mathbf{B})}{\partial z}, \quad (12-74a)$$

$$\mathbf{M} = \mathbf{m} \times \mathbf{B}. \quad (12-74b)$$

Notice that the expression for the force involves *partial derivatives* of the magnetic field intensity as in the expression (11-42) for the force on an electrical dipole placed in an electric field, which means that the force reduces to zero in a uniform magnetic field. The expression for the torque, on the other hand, involves the magnetic field intensity at the location of the dipole regardless of whether the field is uniform or not.

Recall in this connection that, in the case of an electrical dipole, the expression (11-42) reduces to (11-41) (where we have assumed that the electric dipole moment points along the x-axis, for the sake of concreteness) by virtue of the fact the electric field is a conservative one, i.e., can be described in terms of a scalar potential. This, in general, is not the case for a magnetic field, and so the expression (12-74a) for the force on a magnetic dipole cannot, in general, be cast in a form analogous to eq. (11-41).

Problem 12-22

Imagine a small circular coil of radius $r = 0.001$ m and of $N = 100$ turns lying in the x-y plane of a right handed Cartesian co-ordinate system, with a current $i = 0.5$ A flowing in it in the sense

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

of rotation from the positive x- to the positive y-axis. A long straight wire carrying a current $I = 2.0$ A in the positive direction of the z-axis is stretched parallel to the z-axis in the z-x plane at a distance $d = 0.20$ m (measured along the positive direction of the x-axis) from the origin O. Calculate the torque on the circular coil.

Answer to Problem 12-22

HINT: The coil can be looked upon as a magnetic dipole located at the origin O, of dipole moment $m = (\pi r^2)i$ directed along the positive z-axis (check this out; see sec. 12.8.12.1). The magnetic field intensity due to the long straight wire at the location of the dipole is given by $B = \frac{\mu_0 I}{2\pi d}$, where the magnetic field is directed in the negative direction of the y-axis (see sec. 12.8.8.1). The torque on the coil is then given by the expression $\tau = mB$ (see eq. (12-74b)) and is directed along the x-axis. In other words, $\tau = \frac{\mu_0 r^2 i I}{2d}$. Making use of given values, $\tau = 3.14 \times 10^{-10}$ N·m, tending to make the coil rotate about the positive direction of the x-axis in a right handed sense.

12.8.12.4 Energy of a magnetic dipole in a magnetic field

In the case of an electric dipole, it was found that the force F and the torque M on a dipole in an electric field can be expressed as partial derivatives (with negative sign) of a potential energy function U (see sec. 11.6.4), and the work done by the field for a small translation and rotation of the dipole equals the decrease in the potential energy. This is consistent with the principle of conservation of energy since the work performed is equal to the increase in the kinetic energy of the dipole in its rotational and translational motion.

An analogous result holds for a current loop constituting a magnetic dipole. Thus, the potential energy of the dipole in a magnetic field is given by

$$U = -\mathbf{m} \cdot \mathbf{B}. \quad (12-75)$$

However, this is not the energy of the entire system made up of the current loop, the source of EMF supplying the current in the loop, and the source of EMF that supplies the current responsible for the setting up of the magnetic field \mathbf{B} , in which the dipole

is located, where the kinetic energy is, for the time being, disregarded. When we speak of the rotation and translation of a given dipole in a given magnetic field, we have to take into consideration the energy supplied by these sources in keeping all the currents (the current responsible for the dipole moment and the currents setting up the magnetic field) unchanged.

The true energy (U') of the entire system, is related to U as

$$U' = -U = \mathbf{m} \cdot \mathbf{B}. \quad (12-76)$$

In other words, while U can be identified with the potential energy of the dipole looked at in isolation, it does not represent the total magnetic energy of the entire set-up. Since the latter is given by $\mathbf{m} \cdot \mathbf{B}$, the contribution of the sources of EMF in setting up the magnetic field and getting the current loop inserted in it at a given position and in a given orientation must be given by $2\mathbf{m} \cdot \mathbf{B}$.

Looked at from a fundamental point of view, the work done by stationary magnetic fields in any given process has to be zero (see sec 12.8.4). Hence the potential energy of a coil constituting a dipole can increase in a given magnetic field only at the expense of *electrical energy* supplied by the source of EMF maintaining a steady current in the coil.

12.8.13 Magnetic field: comparison with electrostatics

In section 11.7, we saw that an electric field can be described from a geometrical point of view in terms of the lines of force of the field. Analogously, the magnetic field for a given current-carrying circuit or set of circuits can be described geometrically by drawing the lines of force, as in figures 12-32, 12-33, and 12-35 (see sec. 12.8.7). Recall that the tangent at any point on a line of force indicates the direction of the magnetic intensity vector. One common feature to be found in all these configurations of lines of force is that the lines of force are *closed curves* (ignoring exceptional ones like the axial field line for the circular wire).

This is a fundamental feature distinguishing a magnetic field from an electrostatic field,

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

and tells you that the ‘sources’ generating a magnetic field are of a different nature compared to sources generating an electric field. The latter are point charges or small volume elements of charges. Electric field lines *originate from* and *end up on* such point charges or charge elements, while a magnetic line of force, being a closed line, does not have an origin or an end point. A magnetic field is produced by *currents*, which themselves run in closed loops (we are looking at *steady* currents and their magnetic effects for the present). The current lines and the magnetic lines of force wrap around one another, forming loops. Since the sources of the field are the current loops, the field lines have nowhere to originate from or end up at.

There do not exist sources of magnetic fields resembling point charges generating an electric field. One could *imagine* such sources of magnetic field so as to develop the theory of magnetism in complete analogy with electrostatics, and call these imagined sources ‘magnetic poles’. Analogous to positive and negative charges in electrostatics, these imagined magnetic sources might be called ‘north’ and ‘south’ poles respectively. And analogous to electric dipoles (see sec. 11.6), there would then exist magnetic dipoles as well.

Unfortunately, such an imagined description would not be a valid one because such magnetic ‘poles’ do not exist, and all known magnetic fields are generated by moving charges or current loops. However, magnetic fields generated by tiny current loops at a microscopic level, like those in atoms or molecules, have a partial resemblance to electric fields generated by electric dipoles made up of pairs of positive and negative charges. In section 12.8.12 above, such tiny current loops have been identified as *magnetic dipoles*. Indeed, any current loop, large or small, can be looked upon as being made up of such small loops, and a useful description of stationary magnetic fields can be given entirely in terms of magnetic dipoles, with no direct reference to current loops.

As mentioned above, not all magnetic dipoles at the microscopic level can be explained in terms of current loops, i.e., magnetic dipoles may be considered to be more fundamental than current loops as sources of magnetic fields.

It is this fact that accounts for the tradition of using the concept of 'north' and 'south' magnetic poles (analogous to the charges constituting an electric dipole) in describing magnetic fields. However, even though a magnetic dipole resembles an electrical one, the two are really different from a fundamental point of view, as seen from the difference in the ways the fields produced by these dipoles diverge at the locations of the dipoles. Thus, while an electric dipole can be described in terms of a pair of equal and opposite charges, a magnetic dipole cannot, strictly speaking, be described in terms of a pair of opposite magnetic 'poles'.

12.8.14 Currents and magnetic fields: overview

Let us now look back to the theoretical concepts outlined above. I started from the force between two current-carrying wires because the concept of force is already known to us, and it is a good practice to develop new concepts on the basis of already known ones. I then introduced the idea of the magnetic field as an influence generated by a current-carrying wire in the region surrounding the wire by virtue of which another current-carrying wire experiences a force when placed in this region.

This led us to a quantitative definition and measure of magnetic field intensity (see equations (12-47)-(12-51)) in terms of the force experienced by a current-carrying wire placed in a magnetic field. I then completed the circle by giving you the formula for the intensity at any given point due to a current-carrying circuit (see equations (12-53), (12-54)).

This is equivalent to breaking up the formula for the force between two current-carrying wires into two parts - one giving the intensity of the magnetic field generated by any one of the two wires, and the other giving the force experienced by the other wire in this field. Indeed, the formula (12-45) for the force between two small elements of length in the two circuits concerned is obtained simply by combining the formulae (12-53) and (12-47), one for the intensity due to a length element in the first wire at the location of a second length element in the second wire, and the other for the force experienced by the second length element due to this magnetic field.

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

Having at our disposal the basic formula (12-53) for the calculation of the magnetic field intensity, one can then make up a geometric description of magnetic fields in terms of magnetic field lines which, while resembling electric field lines to an extent, differ from the latter in fundamental ways. Finally, I introduced the idea of magnetic dipoles, pointing out that the concept of magnetic poles is not a valid one.

The formula (12-53) involves the units of current and field strength, along with the unit for the constant μ_0 . The latter is analogous to the constant ϵ_0 we encountered in electrostatics. The two constants ϵ_0 and μ_0 are needed in the theory of electricity and magnetism in order to make all the equations used in the theory consistent with one another. One interesting thing about these two constants is that these can be combined to express the value of another fundamental constant of nature, namely the *velocity of electromagnetic waves in vacuum*, which goes to show that there must be some physics relating the two. This is really a reflection of the fact that apparently independent theories and fields of observation are often related to one another at a deeper level.

1. **The magnetic vector potential.**

In comparing electric and magnetic fields, one observes that the electric field produced by stationary charges can be described in terms of an electrical potential, to which the electric field intensity is related as in eq. (11-16). The potential being a scalar function of position, is often more advantageous to work with as compared to the intensity. In the case of a magnetic field produced by steady currents, a corresponding description in terms of a scalar potential does not exist. This distinction between the two fields is once again related to the basic fact that the elementary sources of an electric field are charges, while the elementary sources of a magnetic field are closed current loops, and is reflected in the distinction between electric and magnetic lines of force pointed out in section 12.8.13.

However, one can describe the magnetic field produced by steady currents in terms a *vector potential* instead of the magnetic field intensity, where the latter can be determined once the vector potential is known as a function of position. Though a vector in nature, the vector potential is often of greater advantage in the

description of a static magnetic field.

In the more general case of an *electromagnetic field* where both the electric and magnetic field intensities vary with time, one needs *both* a scalar and a vector potential in describing the field. Once the scalar and the vector potentials are known as a function of position *and* time, one can work out the electric and magnetic field intensities characterizing the field.

2. Having told you about the distinction between static electric and magnetic fields, I have to, in a manner of speaking, make amends by pointing out that, at a deeper level, electric and magnetic fields are *related* to one another. Indeed, static electric and magnetic fields are special instances of more general, time-varying, fields, and may be considered as idealizations since in practice the field intensities vary with time, though the rate of variation may be so slow as to be negligible. As seen in *electromagnetic theory* (see chapter 14), such varying electric and magnetic fields are not independent of each other.

As a simple instance of the interdependence of electric and magnetic fields, consider one or more stationary charges in a certain frame of reference. In a second frame of reference moving with a certain velocity with respect to the first, the charges are no longer stationary, and their motion can be described in terms of a *current density* distributed in space. Such current distributions, in general, give rise to magnetic fields, of which the fields generated by steady currents are particular instances.

Thus, in other words, the categorization in terms of electric and magnetic fields is a *relative* one, and a more general description would be in terms of *electromagnetic fields*, as we will see in chapter 14. In this more general description, the electric and magnetic field components of an electromagnetic field are described in analogous terms, and the description gets altered in a different frame of reference.

In mathematical terms, the electric field intensity for a stationary electric field is a *polar* vector, while the magnetic field intensity of a stationary magnetic field is an *axial* vector (see section 2.6.1 for a brief introduction to these concepts). The mathematical description of an electromagnetic field involves the use of *both*

polar and axial vectors. Together, these make up a *tensor* in *four dimensions*. However, such a level of mathematical description lies outside the scope of the present elementary exposition.

3. Stationary electric and magnetic fields are idealized constructs in the sense that the interaction of charges and currents that they describe are all of the *action-at-a-distance* type (see sec. 3.18 for a brief explanation of the term action-at-a-distance, and the limitations of this concept). As mentioned above, a more realistic description would be in terms of an electromagnetic field where, moreover, one finds that the latter appears as a *dynamical system* in itself, capable of carrying energy, momentum, and angular momentum, and of transferring these to other dynamical systems made up of particles. The electrical and magnetic interactions between particles can thereby be looked upon as ones mediated by the electromagnetic field, where the mediation occurs by means of electromagnetic *waves* propagating in space with the velocity of light in vacuum, compared to which all commonly observed velocities are of an extremely small magnitude. In an approximate description, then, the interactions between particles can be assumed to take place *instantaneously*, accounting for their apparent action-at-distance features.

12.9 Magnetic properties of materials

Up until now, we have had magnetic fields set up by current sources placed *in vacuum*. In reality, however, a magnetic field is produced in some material medium or other. What is of crucial importance here is that the medium itself leaves its stamp on the field that results from given sources. In other words, a given distribution of current sources leads to markedly different magnetic fields in different material media. One expresses this by saying that the various different media differ in their *magnetic properties*. The magnetic properties of materials depend, ultimately, on their *microscopic* constitution.

As a magnetic field is set up in a medium by a given distribution of current sources, the medium itself gets *magnetized*. Looking at any small volume element in the medium, one finds, in general, that it gets endowed with a magnetic *dipole moment*, i.e., in other words, a volume distribution of dipole moment is set up in the material of the medium. Each volume element acting as a dipole, produces a field of its own, and the fields due

to all these dipole sources add up to modify the field that *would have been produced* if the given current sources were placed in vacuum.

For instance, fig. 12-40 depicts a long solenoid carrying a current, with a block B of some given material placed in its interior. In the absence of the block, the field at any given point, located in the region occupied by the block, produced by the solenoid current would have been given by the formula (12-61c) (air is close to vacuum in its magnetic properties). The presence of the block, however, alters things because the block is endowed with a distribution of dipole moments. The field produced by these dipoles gets superposed with the vacuum field to give rise to the resultant field in the block.

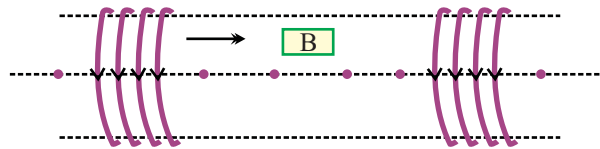


Figure 12-40: A block B made of a given material placed inside a solenoid; illustrating the idea of magnetic properties of a material; the vacuum field produced by the solenoid is represented by the arrow; the field in the material of the block differs from this vacuum field; parts of the solenoid coil are shown while other parts are represented by dots.

With reference to their magnetic properties, material media can be broadly classified into three categories: the *diamagnetic*, the *paramagnetic*, and the *ferromagnetic* ones. The distinctive features of each of these three types of materials relate, in the ultimate analysis, to the way magnetic dipole moments are generated in these at the microscopic level and to the way these are influenced by their environments.

12.9.1 Magnetization in a material body

Considering a small volume element δv around any given point P in a material medium, let the magnetic dipole moment of this element be $\delta \mathbf{m}$. The magnetic moment per unit

volume $\frac{\delta \mathbf{m}}{\delta v}$ is then referred to as the *magnetization* in the medium at that point:

$$\mathbf{M} = \frac{\delta \mathbf{m}}{\delta v}. \quad (12-77)$$

If the current sources producing the field in the material are given, like the solenoid current in fig. 12-40, the magnetization in the material can, in principle, be determined from a detailed consideration of its microscopic constitution. In numerous situations of practical interest, a formula for the magnetization can be worked out, expressing the magnetic property of the material under consideration.

12.9.2 Magnetic susceptibility and magnetic permeability

In this context another classification of magnetic materials is found to be of relevance, namely, the classification into *linear* and *non-linear* ones. Let us consider the linear materials first.

In order to relate the magnetization to the current sources giving rise to the magnetic field intensity in a material, it is useful to introduce a field vector (\mathbf{H}), referred to as the magnetic field strength (as distinct from the field *intensity* \mathbf{B}). For a given distribution of the current sources, termed the *free* currents, the field strength \mathbf{H} is related to these sources in the same way (up to a factor of μ_0^{-1}) as the field intensity \mathbf{B} would have been, *had* the field been produced in vacuum.

1. The term free current distinguishes the current sources of an external nature setting up the magnetic field in a material, from the internal current distribution, of a microscopic origin, that contributes to the resultant field. Thus, for instance, the solenoid current in fig. 12-40 constitutes the free current in respect of the field produced in the material of the block B while the effective current distribution describing the magnetization in the material is referred to as the *bound* current. The resulting field in the material is a consequence of *both* the free and the bound currents.
2. The field \mathbf{H} produced in a medium is related to the free source currents in the same manner as the vacuum- \mathbf{B} (up to a factor), but the two vector fields may

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

in reality be quite different. This relates to the difference in the *boundary conditions* satisfied by the two. Similar considerations apply to the fields \mathbf{D} and \mathbf{E} in electrostatics.

For a linear material, the magnetization \mathbf{M} is found to be *proportional to* \mathbf{H} , where the constant of proportionality is termed the *magnetic susceptibility* (χ_M) of the material:

$$\mathbf{M} = \chi_M \mathbf{H}. \quad (12-78)$$

It is the constant χ_M that characterizes the magnetic property of the material under consideration. The value of this constant can be worked out from microscopic considerations to a good degree of approximation. The *permeability* (μ) of the material is related to the susceptibility as

$$\mu = \mu_0(1 + \chi_M). \quad (12-79)$$

At times, the *relative permeability* (μ_r) of the medium, defined as

$$\mu_r = \frac{\mu}{\mu_0} = 1 + \chi_M, \quad (12-80)$$

is used to describe the magnetic property of a material.

The magnetic field strength \mathbf{H} and the magnetic field intensity \mathbf{B} (also referred to as the ‘flux density’ so as to distinguish it from the field strength) in a medium are related to each other through the permeability as

$$\mathbf{B} = \mu \mathbf{H} = \mu_r \mu_0 \mathbf{H}. \quad (12-81)$$

It has to be emphasized here that the field \mathbf{B} occurring in the above formula *differs* from the vacuum- \mathbf{B} mentioned above.

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

Making use of the permeability μ , one can write down the basic formulae for the magnetic field intensity \mathbf{B} in a linear material due to given distribution of free currents in close analogy with the formulas (12-56) and (12-63) that we wrote down for current sources placed in vacuum:

$$\mathbf{B}(\mathbf{r}) = \frac{\mu}{4\pi} I_{\text{free}} \oint \frac{d\mathbf{l} \times \mathbf{u}}{u^3}, \quad (12-82)$$

$$\oint_C \mathbf{B} \cdot d\mathbf{l} = \mu \sum I_{\text{free}}, \quad (12-83)$$

where the symbols carry the same meanings as already explained, and where I_{free} stands for the free current. Incidentally, equations (12-56) and (12-82) describe the fields set up by a single current loop (constituting the free current in the case of (12-82)), while the formula for the field set up by a number of loops is obtained by making use of the superposition principle.

Strictly speaking, formula (12-82) is valid when the field due to the current loop is produced in an infinitely extended medium; in a material medium with boundaries, bordering on one or more other media, the formula is to be modified by taking into consideration appropriate *boundary conditions*. The working out of the magnetic intensity \mathbf{B} at various points in the medium is then a problem of considerable difficulty.

In terms of the magnetic field strength \mathbf{H} in a material medium, the formula expressing Ampere's circuital law assumes the form

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = \sum I_{\text{free}}, \quad (12-84)$$

As seen from this formula, the unit of magnetic field strength \mathbf{H} is $\text{A}\cdot\text{m}^{-1}$.

A *non-linear* magnetic material is one for which the permeability (μ) itself is a function of the magnetic field strength \mathbf{H} (or of the magnetic field intensity \mathbf{B}), the relation between

\mathbf{B} and \mathbf{H} being then of the form

$$\mathbf{B} = \mu(H)\mathbf{H}. \quad (12-85)$$

Ferromagnetic materials are common examples of non-linear magnetic media. The permeability μ of a ferromagnetic material, apart from being a field-dependent quantity, additionally depends on its prior *history* of magnetization - a phenomenon referred to as *hysteresis*.

Finally, if the field intensity in a medium varies periodically at a high frequency, then \mathbf{B} and \mathbf{H} also vary periodically, but their *phases* may be different. This fact is expressed mathematically by saying that the susceptibility (or, equivalently, the permeability) assumes frequency-dependent *complex* values. This feature characterizes electric fields as well where, in the case of time-varying fields, the electrical susceptibility (or, equivalently, the permittivity) of a medium is expressed mathematically as a frequency-dependent complex quantity.

12.9.2.1 Field variables as space- and time averages

As in the case of electric fields in a dielectric (see sec. 11.10.5.5), the magnetization \mathbf{M} and magnetic field intensity \mathbf{B} in a material medium are actually quantities obtained by an appropriate process of averaging over small spatial distances (of the order of inter-atomic separations) as also over small time intervals (i.e., intervals characterizing atomic and molecular fluctuation processes). It is with these averaged vector fields that the field strength \mathbf{H} in the material is defined. The basic principles underlying the theory of magnetic fields produced by steady currents can then be stated as follows: (a) the surface integral of \mathbf{B} over any and every closed surface imagined in the field is zero, and (b) the line integral of the field strength \mathbf{H} over any closed path equals the *free* current crossing through the loop formed by the closed path.

12.9.3 Dia- and paramagnetism

A detailed theory of the magnetic properties of materials has to be based on a *quantum theoretic* analysis. You will find a brief introduction to quantum theory in chapter 16, but that introduction will be too sketchy to properly address the question of working out the magnetic susceptibilities of materials and of determining the physical factors on which the susceptibilities depend. Instead, what I am going to do here is to indicate briefly only a few relevant results of the theory, and that too in *classical* terms as far as possible. Indeed one can, to a certain extent, make use of classical concepts in describing quantum theoretic results in a kind of hybrid theoretical approach that I will follow here.

Strictly speaking, all magnetic properties of materials are *quantum phenomena* (we will have a brief (and sketchy) introduction to quantum theory in chapter 16). Imagine an assembly of particles, including one or more species of charged particles, where the motions of all these particles take place in accordance with the classical theory. Starting from the system in thermodynamic equilibrium with no net magnetic moment in the absence of a magnetic field, imagine that an external magnetic field is now imposed on it. Assuming that the system with the magnetic field is again in thermodynamic equilibrium, the principles of statistical mechanics imply that it cannot have a magnetic moment in the altered state of equilibrium too. This result, established by Bohr and van Leeuwen, tells us that quantum theory has to enter in an essential way in the explanation of magnetic properties of matter. In the following, we will mostly invoke the classical picture of the motion of charged particles in a magnetic field in explaining the magnetic properties, but quantum principles will play an *implicit* role in the explanation, without being stated and elaborated explicitly. In other words, we will adopt a *semi-classical* approach which is a hybrid one that makes use of the classical way of description involving quantum mechanical objects. The results of such an approach agree with those arrived at in a fully quantum mechanical treatment provided that one is judicious enough in interpreting the intermediate and final results of the theory.

12.9.3.1 Paramagnetism

The atoms of a material are made up of electrons orbiting around the respective nuclei in a set of *stationary orbitals*. The term ‘orbital’ points to the quantum theoretic nature of the state of motion of an electron around the atomic nucleus, where the classical concept of an orbit loses validity. Still, it helps to refer to the quantum states of the electron in terms that have a certain degree of correspondence to the classical orbits. An orbital can be thought of as a tiny current loop, associated with a magnetic moment. Apart from the magnetic moment associated with the orbital by virtue of the motion of the electron around the nucleus, a magnetic moment also results from the intrinsic *spin* of the electron (see chapter 18). In summary, the possible quantum mechanical states of the electron are associated with magnetic dipole moments, depending on the *quantum numbers* characterizing that state (once again, see chapter 18).

The orbitals correspond to certain specific orientations of the angular momentum vector and the associated magnetic moment in space. When the magnetic moments corresponding to all these different orientations are added up for all the electrons in an atom, and then for all the atoms in a macroscopic sample of a material, the resultant magnetic moment component in any given direction is found to be zero in the absence of a magnetic field set up in the medium (see fig. 12-41 where the possible orientations of a typical elementary atomic magnetic moment are shown and where the z-axis is chosen out among all possible directions as the one along which a magnetic field may be imposed as an external influence on the atom). This corresponds to the unmagnetized state of the material under consideration.

While fig. 12-41 depicts the possible orientations of atomic magnetic moments in a material, all lying in a single plane, in reality a more appropriate picture is one where all these moments execute a *precessional* motion around the z-axis (or, for that matter, about any axis chosen in space). In the presence of a magnetic field along the z-axis, the precessional motion about the z-axis continues, but a net magnetic moment results about this axis as explained below.

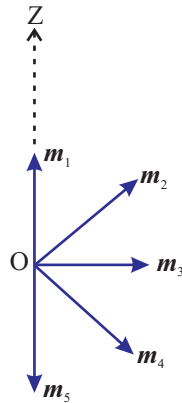


Figure 12-41: Depicting the possible orientations of the resultant magnetic moment of an atom with respect to any given direction in space; five possible orientations are shown for the sake of concreteness; all the orientations are equally likely in a large population of atoms in the absence of a magnetic field; while it appears that there may result a net magnetic moment along the direction of m_3 in the figure, in reality the resultant along each and every direction is zero; this can be understood by imagining all the moments to *precess* uniformly around the z-axis; in the presence of a magnetic field, the precession continues, but the probabilities for the different orientations become unequal.

More precisely, considering the large number of atoms in any volume element of the material, all the different orientations of the magnetic dipole moments turn out to be *equally likely* in the absence of a magnetic field. For instance if m_1, \dots, m_5 represent five possible orientations of the resultant magnetic moment of an atom relative to any given directions in space (the line OZ in fig. 12-41), then out of a large number of atoms, *equal* fractions, on the average, will have their moments oriented along the five directions, with no direction being favored over any other.

Assuming, now, that a magnetic field is set up in the material, some of the orbitals, with their magnetic moments oriented toward the magnetic field become *favored* compared to others with their magnetic moments oriented away from the field. Here is how such a situation comes about.

As one observes from eq. (12-75) in section 12.8.12.4, the energy of magnetic dipole placed in a magnetic field decreases as the angle between the dipole axis and the direction of the field decreases. For instance, out of the five possible orientations of the atomic magnetic dipole moment, the one marked as m_1 in fig. 12-41 will have a lower

energy compared to the other orientations when the atom is placed in a magnetic field along the axis OZ.

Thus, for a large population of atoms, the maximum number will be found to have their dipole moments represented by m_1 while progressively smaller numbers of atoms will be found with moments m_2, \dots, m_5 . In other words, the sum of the components of the magnetic moments along the direction of the magnetic field for the ensemble of atoms under consideration turns out to be non-zero. Put differently, as a magnetic field is set up in a medium made of the material under consideration, any volume element containing a large number of atoms develops a magnetic moment along the direction of the field, i.e., the material is magnetized.

The magnetization so developed depends on the *temperature* of the material because the *probabilities* of an atom to be found in the states of various different energies are determined in accordance with the Boltzmann principle introduced in section 8.14. Indeed, an expression for the susceptibility can be worked out by making use of these probabilities in calculating the magnetization produced in the material due to a magnetic field being set up in it. The expression for the susceptibility is found to be of the general form

$$\chi_M = \frac{C}{T}, \quad (12-86)$$

where C is a constant depending on the material under consideration and T stands for the temperature. This inverse proportionality of the susceptibility with temperature is referred to as *Curie's law* of paramagnetic susceptibility.

In summary, according to the above picture, paramagnetism of a material is due to the individual atoms of the material possessing non-zero magnetic moments. In the absence of a magnetic field, all the possible orientations of the atomic dipoles are equally likely, and there is no resultant dipole moment along any chosen direction. When a magnetic field is set up in the medium, it acts independently on the individual dipoles, tending to orient the dipoles along its own direction. Thus, dipoles with their moments oriented toward the field direction become favored over those with components oriented away

from the field direction, resulting in a net magnetization of the medium.

The electrons in an atom are grouped into closed shells and subshells (see chapter 18), while some atoms may contain incompletely filled shells. The filled shells and subshells in an atom do not contribute to its magnetic dipole moment, i.e., in other words, the resultant magnetic moment of the electrons in such a shell or subshell turns out to be zero. Even in an incompletely filled subshell, electrons are usually paired up to form combinations with zero magnetic moment. Even when an isolated atom possesses a non-zero magnetic moment due to unpaired electrons, the magnetic moment may get canceled when such an atom combines with other atoms to form a molecule. Moreover, when the atoms of a material get together to form a crystalline structure, all the electrons in the unfilled shells become *delocalized*, being shared by all the atoms in the crystal structure (see chapter 19), and the magnetic moments of these delocalized electrons may again get canceled due to pairing of these electrons. There remain only a relatively few unpaired electrons that then contribute to the paramagnetic susceptibility of the material.

In other words, only a relatively few materials with the right type of atomic or molecular compositions and states of aggregation exhibit the paramagnetic property. In the gaseous state, each of the molecular dipoles is acted upon by the applied magnetic field independently of the others since the molecules do not interact appreciably with one another. If the molecules interact more strongly so as to form a solid, then the unpaired electrons belonging to the different molecules again pair up, so that no net magnetic moment remains. There exist a number of crystalline materials on the other hand, where the magnetic moment arises due to electrons in the *inner* shells of the atoms in the crystal. In such solids the inner shell electrons belonging to the different atoms do not interact with one another to form pairs, and there arises a resultant paramagnetic moment. Most of the solids, however, are devoid of the paramagnetic property by virtue of the pairing up of the delocalized electrons. These are, in general, diamagnetic in nature, though in some cases these may have a weak paramagnetic property competing with their diamagnetic nature (see section 12.9.3.2), and the material in question may turn out to be either diamagnetic or paramagnetic as a result of this competition.

As a result of an externally imposed magnetic field in a paramagnetic material tending to align the elementary magnetic dipoles in its own direction, the susceptibility of a paramagnetic material is positive, and its relative permeability satisfies $\mu_r > 1$. When placed in a non-uniform magnetic field, a sample of a paramagnetic material gets drawn from a region of weaker field to one where the field is stronger since this results in a lower magnetic energy of the system. As we will see below, this contrasts with what happens for a diamagnetic material.

The magnetic dipole moment of an atom or a molecule is typically of the order of a *Bohr magneton* (μ_B), where the latter is the quantum mechanical unit of magnetic moment given by the expression

$$\mu_B = \frac{eh}{4\pi m_e}, \quad (12-87)$$

where e stands for the electronic charge, m_e for the mass of the electron, and h denotes the *Planck constant* (6.626×10^{-34} J·s).

The paramagnetic susceptibility of a material works out to a value of the order of

$$\chi_M \sim \frac{N\mu_0\mu_B^2}{k_B T}, \quad (12-88)$$

where k_B stands for the Boltzmann constant (1.38×10^{-23} J·K⁻¹), and N for the number density of the elementary magnetic moments.

12.9.3.2 Diamagnetism

While paramagnetism and ferromagnetism (see sec. 12.9.4) are magnetic properties exhibited only by certain materials meeting specific requirements, diamagnetism is a *universal* property of materials. Indeed, diamagnetism is an intrinsic property of atoms (and molecules) or, to be more precise, of each quantum mechanical electron orbital in the atoms. In a paramagnetic or a ferromagnetic material the diamagnetic features are eclipsed by the para- or the ferromagnetic features, as the case may be. All the remaining materials then appear as diamagnetic ones.

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

One way to explain diamagnetism is to say that it is *a consequence of the phenomenon of electromagnetic induction* (see section 13.2) *operating at a microscopic level*. Imagine, for the moment, a closed circular loop (fig. 12-42) of current (I) caused by a charged particle of charge q and mass M moving along the circle with an angular velocity ω . The current I is then given by the expression

$$I = \frac{q\omega}{2\pi}. \quad (12-89)$$

The associated magnetic moment (related by the right hand rule to the direction of the current along the bent arrow in the figure, assuming q to be positive, for the sake of concreteness) is given by

$$m = \pi r^2 I = \frac{q\omega r^2}{2}. \quad (12-90)$$

Suppose now that a magnetic field with field intensity B is set up perpendicular to the plane of the circle (in the upward direction in fig. 12-42) in time τ . This involves a change in the magnetic flux through the current loop, the rate of change being $\frac{\pi r^2 B}{\tau}$. As a consequence, an EMF will be induced in the loop in a direction opposite to the one related to the direction of the field by the right hand rule. This EMF results in an electric field intensity E appearing at every point on the circular loop acting tangentially, where

$$E = \frac{1}{2\pi r} \frac{\pi r^2 B}{\tau}. \quad (12-91)$$

The electric field applies a force $F = Eq$ on the circulating charge operating during the interval τ , in the direction *opposite* to the one indicated by the bent arrow in the figure. The resulting torque causes a change in the angular momentum (\mathbf{L}) of the particle given by

$$|\Delta \mathbf{L}| = qEr\tau = \frac{qBr^2}{2}, \quad (12-92)$$

and a corresponding change in the magnetic moment associated with the circulating

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

current loop (refer to equation (12-73)),

$$\Delta \mathbf{m} = -\frac{q^2 r^2 \mathbf{B}}{4M}. \quad (12-93)$$

The direction in which the magnetic moment changes is opposite to the direction of the impressed magnetic field *regardless of the direction of motion of the charge in the circular orbit and of the sign of the charge* (check this out).

In other words, a magnetic moment is induced in a current loop due to the setting up of a magnetic field of intensity B , in a direction opposite to the field. This is the essential phenomenon involved in diamagnetism.

Considering now the atomic current loops in a material medium, one can conclude that, as a magnetic field is set up in it, a magnetic moment will be induced in the material in a direction opposite to the field, the induced moment being proportional to the average of the squared radii of the electronic orbits in the atoms, where, moreover one has to replace q and M with the electronic charge e and mass m_e respectively. Making use of eq. (12-93), one can work out an order of magnitude expression for the diamagnetic susceptibility of a material, where the susceptibility is now *negative* because of the direction of the induced magnetic moment being opposite to the magnetic field.

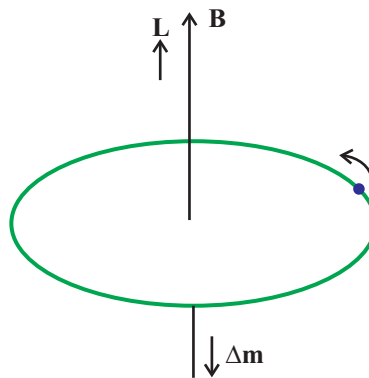


Figure 12-42: Induced EMF in a current loop - explanation of diamagnetism; a charge q is shown circulating in an orbit, where the orbital angular momentum is L ; as a magnetic field B is introduced in the direction shown, an induced magnetic moment Δm is produced in an opposite direction.

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

In the above example, the current loop associated with the circulating charge possesses a magnetic moment to start with, which then gets changed by Δm as the magnetic field is set up. In general, the electron orbitals in an atom or a molecule are characterized by non-zero magnetic moments where, in many cases, the magnetic moments of all the orbitals in an atom or a molecule cancel one another. These correspond to the materials that exhibit the diamagnetic property. Though the magnetic moment is zero in the absence of the magnetic field, an induced magnetization appears as the field is imposed, in virtue of the diamagnetic effect on each individual electron orbital, the corresponding susceptibility being *negative* because of the direction of the induced magnetization being opposite to the magnetic field.

If the net magnetic moment of each atom or molecule is zero, then there is no question of a resultant magnetic moment appearing in an assembly of atoms by virtue of an imposed magnetic field by the mechanism outlined in sec. 12.9.3.1, i.e., in other words, the material cannot, in the commonly understood sense of the term, exhibit the paramagnetic property.

The above derivation constitutes an explanation of diamagnetism as a universal magnetic property of materials, in *classical* terms. Strictly speaking, one has to make use of quantum concepts in order to arrive at a satisfactory theory of diamagnetism. In a quantum theoretic derivation, the expression for the induced magnetic moment for any given atomic orbital works out to be the *same* as in eq. (12-93), where r^2 is to be replaced with $\langle r^2 \rangle$, the quantum mechanical average value of the squared distance of the electron from the nucleus. In such a quantum theoretic derivation, the angular momentum is to be interpreted in a *generalized* sense, where the generalized angular momentum depends on the magnetic field intensity B . This generalized angular momentum can take up only a discrete set of values in quantum theory, as opposed to a continuously distributed set of values in the classical theory. As the magnetic field is set up, the generalized angular momentum remains constant, while the mechanical angular momentum ($\mathbf{r} \times \mathbf{p}$) changes, with an attendant change in the magnetic moment given by eq.(12-93).

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

The diamagnetic susceptibility worked out from eq. (12-93) is found to be *independent* of the temperature, distinguishing it from the paramagnetic susceptibility which varies inversely as the temperature in accordance with Curie's law (eq. (12-86)).

One other feature distinguishing diamagnetism from paramagnetism is the negative sign of the susceptibility χ_M , in consequence to which a sample of a diamagnetic material, when placed in a non-uniform magnetic field, tends to be displaced from a region of higher field intensity to a region of a lower one.

Since diamagnetism is a universal magnetic property of materials, a paramagnetic or a ferromagnetic (see sec. 12.9.4) material also develops an induced magnetic moment in a magnetic field given by an expression of the form (12-93), though the resulting magnetization is masked by the magnetization developed *along* the direction of the field by virtue of the paramagnetic or the ferromagnetic property. In general, the susceptibility of a paramagnetic material is larger than that of a diamagnetic one by one or two orders of magnitude, while ferromagnetic susceptibilities are larger by several orders of magnitude.

Conducting materials are generally diamagnetic where this diamagnetism is related to the existence of a pool of *free electrons* serving as electrical carriers in a conductor. A few conducting materials, on the other hand, are endowed with the property of ferromagnetism that completely masks the diamagnetic features. Superconducting materials, characterized by zero resistivity, behave as *perfect diamagnets*.

The diamagnetic property of conductors, by virtue of which a conducting material is repelled from a region of a strong magnetic field to one of a weak field, is made use of in *diamagnetic levitation*.

Diamagnetism is a tricky subject, where one may get trapped in pitfalls of reasoning. One paper [12] I can refer you to is by S.L. O'Dell and R.K.P. Zia, entitled 'Classical and semi-classical diamagnetism: A critique of treatment in elementary texts' (American Journal of Physics, vol. 54, p 32-35, 1986).

Problem 12-23

A small piece of a certain diamagnetic metal, of volume $\delta V = 2.1 \times 10^{-9} \text{m}^3$, when inserted inside a long solenoid of $N_0 = 10,000$ turns per meter gets magnetized, under the influence of a magnetizing current of $I = 2\text{A}$ in the solenoid, the dipole moment produced in the sample being $\delta m = 4.1 \times 10^{-10} \text{A}\cdot\text{m}^2$; assuming that the magnetic field strength H (along the axis of the solenoid) in the sample is $\frac{1}{\mu_0}$ times the flux density B that would be produced in its absence, obtain the magnetic susceptibility of the material.

Answer to Problem 12-23

SOLUTION: Making use of the given relation between the field strength H , directed along the axis of the solenoid, and the magnetic flux density B that would be produced in the absence of the sample, and of the formula (12-67), one obtains $H = N_0 I$. The magnetic susceptibility is then obtained from formulae (12-78) and (12-79) as $\chi_M = -\frac{1}{H} \frac{\delta m}{\delta V} = -\frac{4.1 \times 10^{-10}}{2.1 \times 10^{-9} \times 10^4 \times 2} = -9.7 \times 10^{-6}$, where the negative sign appears due to the fact that the material is a diamagnetic one, for which the magnetic moment is produced in a direction opposite to the direction of the magnetic field strength.

12.9.4 Ferromagnetism

Ferromagnetism is a very special magnetic property exhibited by certain specific materials, including iron, the most familiar of such materials. Our familiarity with the ferromagnetic behavior of iron often stands in the way of realizing how remarkable this behavior is when looked at in comparison with the magnetic properties of most other materials, which are either diamagnetic or paramagnetic.

12.9.4.1 Spontaneous magnetization

The most remarkable magnetic property of a ferromagnet relates to the phenomenon of *spontaneous magnetization*. While in a paramagnetic or a diamagnetic material, the magnetization develops in response to a magnetic field set up in the material by external (or *free*) current sources, the magnetization in a ferromagnetic material develops spontaneously, even *without* a magnetic field set up in it. However, such spontaneous

magnetization is commonly not observed since it is developed only in small regions called *domains* in the material, where the magnetizations in the various different domains cancel one another and no net magnetization shows up in it. Once again, a field is to be set up in the material for a net magnetization to be exhibited.

Other curious instances of ferromagnetic behavior relate to the phenomena of *hysteresis* and *residual magnetism*, which we will have a look at in sections 12.9.4.3 and 12.9.4.4.

Finally, and equally remarkably, a ferromagnetic material undergoes a *phase transition* at a certain *critical temperature*, when it loses its spontaneous magnetization and behaves like a paramagnetic material.

The explanation of spontaneous magnetization, the most basic feature of ferromagnetic behavior, is to be found in the crystalline structure of a ferromagnetic material where the lattice sites in the crystal are occupied by what are referred to as ferromagnetic *ions*. These ions possess unpaired electrons, where the spin of the unpaired electron in an ion can have any one of two different orientations with respect to any chosen direction in space. These are termed the spin-up and spin-down orientations, corresponding to which the magnetic moment of the electron along the chosen direction can be either $-\mu_B$ or $+\mu_B$, where μ_B stands for the Bohr magneton introduced in sec. 12.9.3.1 (eq. (12-87)).

The spins of the unpaired electrons in the ferromagnetic ions located at the lattice sites of the crystalline structure are not independent of one another, but are correlated by means of interactions between the electrons. These interactions are not magnetic but are primarily *electrostatic* in nature, the latter being much stronger compared to the former. These electrostatic interactions between the electrons of neighboring ions in the lattice are *spin-dependent* due to a constraint imposed by the fact that the electrons are *fermions* (see section 18.8.9.2 for a brief introduction; the spin dependence of the electron-electron electrostatic interaction will again be encountered in the context of atomic energy levels in sec. 18.5.2.3). Ordinarily, in the case of neighboring atoms interacting by means of shared electrons, as in a covalent bond, the spin-dependence of the interaction results in the electron spins being oriented in an *anti-parallel* configura-

tion, i.e., with one of the electrons in the spin-up and the other in the spin-down state (refer to sec. 18.9.1.2). However, for neighboring ions in a ferromagnetic material the crystalline environment of the ions turns out to have a crucial role in that it results in the *parallel* configuration being favored over the anti-parallel one.

In other words, due to the special crystalline environment of the ions in the crystal structure of a ferromagnetic material, the favored configuration of electrons in neighboring ions happens to be the one with their spins (and hence their magnetic moments) being parallel to each other. This, in turn, leads to a *co-operative* interaction where *all* the spins in the lattice tend to be oriented parallel to one another (see fig. 12-43), resulting in the entire system made up of the ions being endowed with a magnetization. For instance, with N number of spins per unit volume oriented parallel to one another, the magnetization developed will be $N\mu_B$. This is the spontaneous magnetization characterizing a ferromagnetic material since it is developed without requiring a magnetic field to be set up in it by external means.

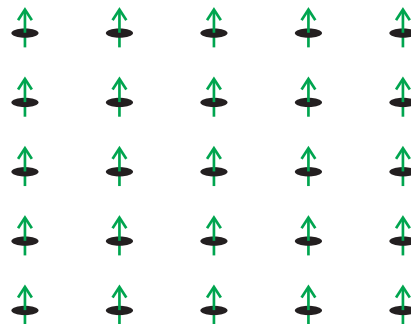


Figure 12-43: Depicting spontaneous magnetization by co-operative alignment of all the spins in a crystal structure; the crystal structure is made up of a regular arrangement of ions while the electrons with their spins aligned in a common direction are represented by the small arrows.

12.9.4.2 Magnetic domains

Two questions are to be addressed at this stage. The first of these is the following: if the spins get aligned all by themselves, why does a piece of iron not ordinarily exhibit a magnetization even without an externally impressed magnetic field being set up in it? The answer lies in the fact that the spins do get aligned but only in relatively small

blocks in the material called *domains*. The entire volume of a ferromagnetic material is filled up with such domains, where each domain is spontaneously magnetized by the co-operative interaction among the spins in it. However, the direction in which the spontaneous magnetization develops in one domain is ordinarily not correlated with the direction of magnetization in any other domain so that, the magnetizations of all the domains taken together cancel one another.

The reason why the spontaneous magnetization is developed in small domains rather than in the entire volume of the ferromagnetic material is that the magnetization of the entire volume is associated with a rather steep energy cost, which is avoided by domain formation. A ferromagnetic crystal possesses one or more *favoured* directions in it such that the magnetization in a domain tends to occur in a direction parallel to that, pointing either one way or the other along the favored direction (see fig. 12-44). The magnetizations in neighbouring domains being directed at random along one or the other of the two opposite directions, the material as a whole exhibits no net magnetization.

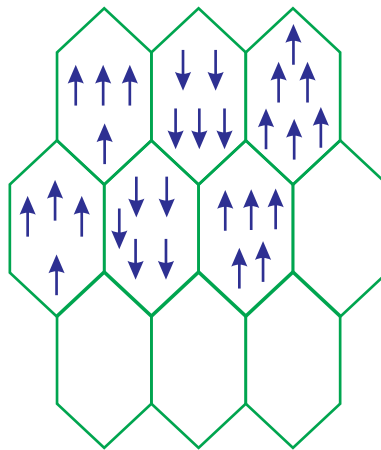


Figure 12-44: Spontaneously magnetized domains in a ferromagnetic material; each domain is a small volume within the material, in which all the spins (associated with the ferromagnetic ions) are aligned either one way or the other (either up or down in the figure) following a certain favored direction (the vertical direction in the figure) within the crystal structure; the spontaneous magnetizations in the various different domains cancel one another unless an external magnetic field favours one of the two opposite magnetization directions.

The second question that comes up now is, how do *permanent magnets* come to possess a magnetization even in the absence of a magnetizing field in spite of the magnetizations of the domains canceling one another? The answer to this question relates to the phenomenon of *hysteresis*.

12.9.4.3 The magnetization curve: hysteresis

Imagine, in fig. 12-40, a sample of a ferromagnetic material placed inside a solenoid that sets up a magnetic field in the material, where the strength of the magnetic field can be varied by varying the current (I) through the solenoid winding. The magnetic field results in the development of an magnetization in the material as a whole. One can experimentally determine the magnetization (M) developed in the material for any given value of the current I , and then plot graphically the variation of M as a function of I . The resulting graph looks somewhat like the one shown in figure 12-45.

Starting with the sample in a demagnetized condition, with no magnetizing current in the solenoid, as the current is made to increase the magnetization increases linearly along the portion OA of the graph. What happens here is a process of expansion of the domains with their magnetization aligned to a greater degree with the magnetic field (the so-called 'favored' domains), at the expense of the other domains. This is a *reversible* process, and the magnetization is found to decrease along AO if the current is decreased. However, as the magnetization proceeds beyond a certain stage, a further expansion of the favored domains is hindered by *imperfections* in the crystalline structure. The magnetization now increases with the current in an *irreversible* manner as the domains expand by snapping past the imperfections. If the magnetizing current is now made to decrease, the domains do not contract in a reverse sequence since the contraction is again hindered by the imperfections. Every time a domain wall is forced past an imperfection, a certain amount of energy is consumed in the process, which gets dissipated in the material in the form of heat.

With increasing current, the magnetization ultimately reaches a saturation value at B with most of the domains now turned around toward the direction of the field. If now

the current is decreased, the magnetization decreases following a different course along BC because of the irreversibility of the magnetization process, and as the magnetizing current (and hence the magnetizing field H) is made to decrease to zero value, a magnetization, corresponding to the segment OC *remains* in the sample. If the current is now reversed and then made to increase in magnitude, the portion CD of the graph is traced out, and finally, a cycle is completed along DEB as the magnetizing current is taken through a process of cyclic change. The corresponding closed curve in the graph describing the cyclic magnetization process is termed a *hysteresis loop*

For any given magnetizing current, say, I_1 in the figure, one finds several possible values of the magnetization (M_1, M_2, M_3), depending on the prior sequence of values through which this current is established. In other words, the degree of magnetization depends on the *history* of the process. This is the phenomenon of *hysteresis*, which arises due to the irreversibility of the magnetization sequence. As the current is made to pass through a complete cycle of values, a complete loop of the magnetization curve is traced out, and a certain amount of energy has to be expended in the process which appears as heat in the material.

12.9.4.4 Residual magnetism: permanent magnets

Suppose that a ferromagnetic sample is magnetized up to saturation (point B in the graph of fig. 12-45) and then withdrawn from the magnetizing field. In the experimental situation considered above, this means that the magnetizing current is switched off to zero or else the sample is taken out from the interior of the solenoid. As indicated above, this does not result in the sample being demagnetized since the demagnetization does not follow the prior process of magnetization in reversed order. Rather, the domain configuration tends to be retained in the sample, and one reaches the point C in the graph of fig. 12-45 where the sample retains a non-zero magnetization (M_{residual}) represented by the segment OC in the graph. This is referred to as the *residual magnetization* in the material.

Various alloys of iron, nickel, and cobalt (all of which are ferromagnetic materials) are

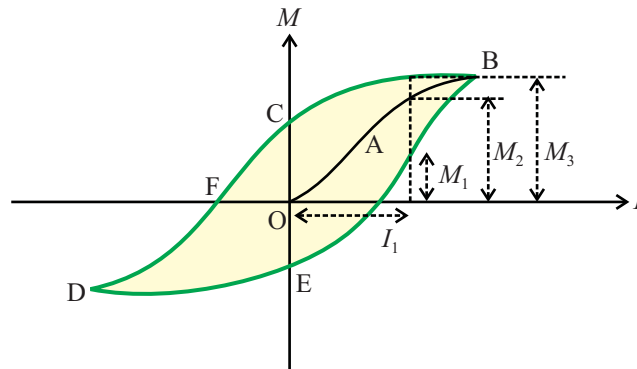


Figure 12-45: Magnetization curve of a ferromagnetic material, plotted with the magnetization M developed in a sample, against the magnetizing current I in the electromagnet used in the magnetization process; starting from the demagnetized condition, M increases linearly with I along OA in a reversible manner; the curve from A to B is nonlinear where the variation of M with I is irreversible in nature; B represents saturation of the magnetization; as I is decreased back to zero, M decreases along BC to a non-zero value (residual magnetism); as I is reversed and increased in magnitude, the magnetization decreases to zero at a non-zero reversed current (point F, corresponding to a value of H termed the coercive field), and then attains saturation in the reversed direction (point D); as I is then increased, a symmetric branch (DEB) is traced out; the phenomenon of hysteresis is indicated by three different values of the magnetization (M_1 , M_2 , M_3), depending on the history of the magnetization process, for the same magnetizing current (I_1).

characterized by various different values of residual magnetization. Materials with a high value of the residual magnetization are used to produce *permanent magnets*. A permanent magnet is produced by placing a sample of ferromagnetic material, characterized by a large residual magnetism, in a strong magnetic field produced by an electromagnet as in fig. 12-46, and then switching off the magnetizing current. The residual magnetism remains in the material for a long time when the sample acts as a magnet, i.e., produces a magnetic field even without any free current source such as a current-carrying coil being used for the purpose. It is the dipoles at the microscopic level that produce the field by virtue of their ordered arrangement in the crystalline lattice.

Incidentally, an *electromagnet* is a set-up shown schematically in fig. 12-46, consisting of a winding (usually made of a thick wire designed to carry large currents) on an iron frame, with a small air gap (A) in between two *pole-pieces* (P, P'), the latter being parts of the frame. On passing a large current through the winding, a strong magnetic field is set up in the air gap between the pole-pieces on account of the permeability of the material of the frame (the 'core') being high. The magnetic field disappears with the

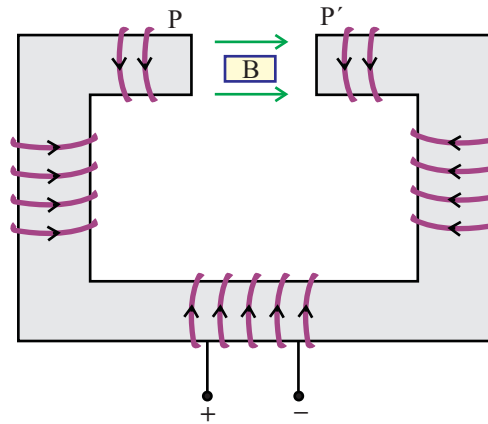


Figure 12-46: Production of a permanent magnet using an electromagnet; the latter is made of a suitably shaped magnetic frame with a coil wound on it; as a current is passed through the coil, the frame is magnetized and a strong magnetic field (represented by arrows) is produced between the *pole pieces* (P,P'); a sample (B) of a suitable magnetic material placed between the pole pieces is thereby magnetized, and retains part of its magnetism even when the magnetizing current is switched off.

current being switched off because the composition of the core material is so chosen as to make its residual magnetism very low.

Pieces of certain ores of iron are seen to behave as permanent magnets. Their magnetization is likely to have developed in the past owing to these ores having been subjected to some strong magnetic field, possibly of geomagnetic origin (see section 12.10). Needle-shaped permanent magnets are used in navigation as direction-pointers. A freely suspended rectangular or needle-shaped permanent magnet orients itself along the earth's magnetic field (see sec. 12.10) where the field lines run approximately in a south-north direction. The end of the magnet pointing toward the north is termed the 'north pole' of the magnet, while the other end, pointing south is referred to as the 'south pole'. These terms are used despite the fact that isolated magnetic poles have not been observed.

Looking at fig. 12-45 one observes that, after magnetizing a ferromagnetic material, say, up to the point B, it takes a *reverse* current of a certain magnitude to demagnetize it (the point F between C and D on the loop corresponding to $M = 0$). The reverse field strength (H) corresponding to this point is referred to as the *coercive field*. Materials needing a comparatively strong coercive field are referred to as *hard* ones, while a weak coercive

field corresponds to a *soft* material. Properties of magnetic materials like the residual magnetism and the coercive field are made use of in *magnetic recording*, as also in the reading and erasing of magnetically written information.

12.9.4.5 Transition to paramagnetic behaviour

The co-operative interaction of the spins in a ferromagnetic material is an *ordered* configuration where, knowing the *macroscopic* state of the system, characterized by the magnetization M (in a domain which is a small but macroscopic volume in the material), one can predict with reasonable certainty the *microscopic* states of the individual spins. Such an ordered configuration of the system of spins corresponds to the state of lowest interaction energy of the spins.

However, the lowering of energy is not the only criterion governing the macroscopic state of a system. Another criterion, of a competing nature, is the increase of *entropy* or *disorder* of the system (refer to sec. 8.14 for a brief introduction to the concept of entropy). For a system of given volume, held at any given temperature T , a compromise between the minimization of energy (the internal energy in the thermodynamic description) and maximization of entropy is realized by way of the *free energy* of the system being a minimum at the state of equilibrium, where the free energy is a thermodynamic variable like the internal energy and the entropy.

At low temperatures the minimization of free energy effectively reduces to the minimization of the internal energy, and the ordered configuration of spins corresponding to the spontaneously magnetized, or ferromagnetic, state prevails. At high temperatures, on the other hand, the minimization of internal energy ceases to be the determining factor, and the equilibrium state is effectively determined by the principle of increase of disorder (refer to section 8.14.1), when the material loses its spontaneous magnetization and remains demagnetized unless a magnetizing field is set up in it.

In between, there exists a certain *critical* temperature (T_C) where the material makes a transition from the ferromagnetic to paramagnetic behavior. Above the critical temper-

ature, the variation of the susceptibility of the material is of the form

$$\chi_M = \frac{C}{T - T_C}, \quad (12-94)$$

where C is a constant for the material under consideration. As this formula shows, the variation of susceptibility with temperature reduces to the Curie law given by eq. (12-86) for $T \gg T_C$, while, as T is made to approach T_C , the susceptibility diverges, corresponding to the acquisition of spontaneous magnetization.

12.10 The earth as a magnet: geomagnetism

It has been known for long that the earth itself behaves as a huge magnet. The earth's magnetic field is constantly being mapped and monitored, and much is now known on the *past history* of the earth's magnetism as well. Analogous to the earth behaving as a magnet, the sun and other stars have also been found to be endowed with magnetic properties, with a number of common features between the magnetic behavior of all these celestial bodies.

The magnetic field intensity (B) at any location near the earth's surface is commonly described in terms of three parameters - the *horizontal intensity* (B_H), the *dip* (θ), and the *declination* (δ). Consider two vertical planes through the point (P) under consideration (fig. 12-47). One of these is the geographical meridian plane (i.e., a vertical plane containing a horizontal segment (PN) of the longitude circle, connecting the two geographical poles, passing through P) while the other contains the magnetic field intensity vector (B), the direction of the latter being indicated by a freely suspended magnetic needle.

The angle between these two planes is defined as the declination at the point of observation P. The angle between the horizontal line (PM) in the second of the above two planes (referred to as the magnetic meridian plane) and the direction of B is the angle of dip, while the component of B along PM gives the horizontal intensity at P. Determining these quantities at various points throughout the surface of the globe gives a mapping

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

of the earth's magnetic lines of force. The earth's magnetic field has also been charted at different altitudes above the surface of the earth. Evidently, these definitions imply

$$B_H = |\mathbf{B}| \cos \theta. \quad (12-95)$$

The declination at any given location may be described as so many degrees *east* or *west*, depending on whether the north pole of a freely suspended needle is deviated from the geographical meridian to the east or to the west. Similarly, the dip is described as so many degrees *north* or south depending on whether the north pole of the freely suspended needle points downward or upward in the magnetic meridian.

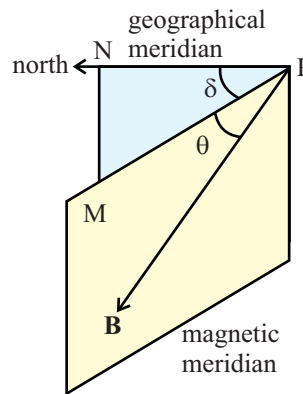


Figure 12-47: Illustrating the three magnetic elements at any given point P; the geographical meridian and the magnetic meridian are shown, being inclined to each other at an angle δ , the declination at the point under consideration; PN and PM are horizontal lines in the two meridian planes; the direction of the magnetic field intensity \mathbf{B} is shown; the component of \mathbf{B} along PM is the horizontal intensity (B_H); the angle of dip (θ) is also shown.

Some of the major features of the variation of the terrestrial magnetic field in time and space are as follows: (a) the field looks approximately like that of a *magnetic dipole* placed at the center of the earth, where the dipole axis is inclined at a small angle with the south-north axis of rotation of the earth (see fig. 12-48); (b) the magnetic field remains nearly constant over large periods of time, but nevertheless has a *slow variation*, where the variation assumes appreciable proportions over thousands of years; (c) the dipole axis *flips over*, i.e., changes direction by nearly 180 degrees, at intervals of about 200,000 years; (d) there is a certain small *multipole* component mixed with the dipole

field.

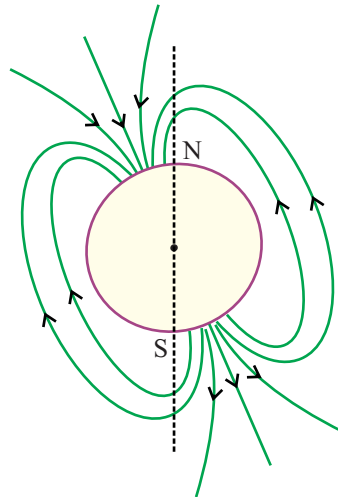


Figure 12-48: A schematic depiction of earth's magnetic field, represented by a set of lines of force, similar to the field of a magnetic dipole; the dipole axis is inclined at an angle to the south-north rotation axis of the earth.

The origin of the earth's magnetism can be explained in terms of what is referred to as the *geomagnetic dynamo* theory, where a complicated set of interdependent events in the earth's core is believed to result in a *self-sustained* dynamic structure involving fluid flow and the production of a magnetic field.

Reduced to bare essentials, the geomagnetic dynamo theory involves the following: The central region of the earth's core, known as the *inner* core is a hot solidified mass at a temperature of the surface of the sun, surrounded by an *outer* fluid core, where both the inner and outer cores are rich in iron. A slow process of solidification of the outer core fluid takes place at the inner core boundary with a release of latent heat, which maintains a convective flow of the fluid. While the relatively heavier iron-rich components of the outer core fluid solidify and get deposited on the inner core, the lighter components remain in the liquid phase and rise towards the earth's surface in a buoyancy-driven flow. The temperature- and buoyancy-driven convective flow is helical in nature. Assuming that the flow takes place in a magnetic field, a current is generated in the fluid, which is an electrically conducting medium, due to a motional EMF being set up in it

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

(see section 13.2.2.2). This current, in turn, produces a magnetic field of its own, which regenerates and adds to the field that generated the current to start with. To complete the story, the Lorentz force exerted by the magnetic field goes to reinforce the current.

It has been established through numerical computations that the entire set of processes outlined above can run in a self-sustaining manner provided the convective flow in the outer core crosses a certain threshold level, where the magnetic field is generated and sustained by the fluid flow in the system itself. A detailed description of the multipole nature of the magnetic field and the explanation of polarity reversal requires that a number of other factors be included in the theory, resulting in a complicated set of mathematical equations. A crucial role in the interdependent set of processes is played by the rotation of the earth, which makes the earth-bound frame of reference a *non-inertial* frame (see section 3.10.3 and 3.21 for an introduction to non-inertial frames, where examples are to be found of *pseudo-forces* that can arise in such a frame), in which the centrifugal and Coriolis forces affect the motion of the fluid in the outer core.

While a complete and conclusive explanation of the earth's magnetic field including all its characteristic features is yet to emerge, the geomagnetic dynamo theory appears to provide a consistent theoretical framework for the purpose.

12.11 The chemical effect of current

12.11.1 Electrolytes and electrolysis

An electrolyte is a compound in which charged atoms or groups of atoms are held together by electrostatic forces, resulting in its ability to conduct electricity, primarily in an aqueous solution. Conduction of current by an electrolyte is, however, also possible in non-aqueous medium, such as through a molten electrolyte or, in some instances, in a gelatinous or even a solid phase. Molten metals, through which electrical conduction occurs by means of electrons rather than ions, are not regarded as electrolytes.

In an aqueous solution of an electrolyte, the forces holding the ions together become

CHAPTER 12. ELECTRICITY I: STEADY CURRENTS AND THEIR MAGNETIC EFFECTS

weaker, causing the ions to drift apart when a pair of electrodes are inserted in the solution, with a potential difference applied between the electrodes.

As mentioned in sec. 11.10.5.2, the electric field intensity in a dielectric medium due to charges setting up an electric field gets reduced compared to the intensity that would result in the absence of the dielectric, by a factor of ϵ_r , the relative permittivity of the dielectric. This reduction in field intensity due to the polarization produced in the dielectric implies a corresponding reduction in the Coulomb force between charges placed in the dielectric, the reduction in the force being once again by a factor of ϵ_r . The relative permittivity of water is large compared to other liquids due to the dipolar nature of the water molecules. This results in a considerable loosening of the ionic binding forces in an aqueous medium. Hence when an electric field is set up in the aqueous medium containing an electrolyte in solution, the positive and negative ions drift apart in opposite directions due to the force exerted by the field. Such an electric field is produced when a potential difference is applied between a pair of electrodes dipped in the electrolyte solution.

This phenomenon of separation of the ions, held together by electrostatic forces in an electrolyte, resulting from the setting up of an electric field in the electrolyte, is referred to as *electrolysis*.

An ion may be positively or negatively charged by way of one or more electrons having been removed from or added to a neutral atom or a group of atoms. Thus, the magnitude of the charge on an ion is an integral multiple of the electronic charge ($e = 1.6 \times 10^{-19}$ C) where a neutral molecule of an electrolyte is made up of a positively and a negatively charged ion having the same magnitude of charge. Calling this magnitude ne , ($n = 1, 2, \dots$), N number of ions of either type make up a charge of magnitude Nne . Here, the integer n stands for the deficit or excess of the number of electrons in the ion, and is commonly referred to as its *valence number*.

Supposing that ν mol of the electrolyte are electrolysed, where each mole gives rise to N_A number of ions ($N_A = \text{Avogadro number} = 6.02 \times 10^{23}$), one will have $N = \nu N_A$, and

thus the magnitude of the total charge of either species of ions will be $q = n\nu eN_A$. In this expression, eN_A stands for the magnitude of the charge of one mole of electrons, and is referred to as the *Faraday constant*, which is used as a unit of charge and is denoted by the symbol F , where $1 F = 9.65 \times 10^4 \text{ C}$.

Thus, in summary, the electrolysis of ν mole of an electrolyte causes a total charge of magnitude νneN_A to migrate towards each of the two electrodes, where the positive ions move towards the negative electrode (the cathode) and the negative ions migrate towards the positive electrode (anode; refer to sec. 12.1.2.4).

On the other hand, the *mass* of the ionic species carrying this amount of charge is given by $M = \nu N_A m_0 A_r$, where m_0 stands for the mass corresponding to one atomic mass unit ($\frac{1}{12}$ th of the mass of a carbon-12 atom, $\approx 1.66 \times 10^{-27} \text{ kg}$) and A_r for the relative ionic mass of the ion under consideration (i.e., the mass of the ion relative to $\frac{1}{12}$ times the mass of a carbon-12 atom)

12.11.2 Faraday's laws of electrolysis

Supposing that the process of break-up of ν mol of the electrolyte molecules is effected by the passage of a current I , for a time t , i.e., by the passage of a charge $Q = It$, one has, from above,

$$Q = \nu neN_A, \quad M = \nu N_A m_0 A_r, \quad (12-96)$$

i.e.,

$$M = \frac{m_0 A_r}{ne} Q. \quad (12-97)$$

The quantity

$$z = \frac{m_0 A_r}{ne}, \quad (12-98)$$

gives the ratio of the mass and the charge of an ion, and is referred to as its electro-

chemical equivalent. In this expression for the electrochemical equivalent z , m_0 and e are universal constants while n , the valence number, and A_r , the relative ionic mass, depend on the ion under consideration. The ratio $\frac{A_r}{n}$ is commonly referred to as the chemical equivalent of the ion. Thus, the electrochemical equivalent and the chemical equivalent are proportional to each other.

The relations (12-97) and (12-98) form the basis of *Faraday's laws of electrolysis*: (a) the mass of any particular ionic species liberated at an electrode is proportional to the amount of charge (Q) passed through an electrolyte, the constant of proportionality being the electrochemical equivalent of the ion; (b) the masses of a number of different ionic species liberated by the passage of a given quantity of charge are in proportion to their chemical equivalents.

12.12 Thermoelectric effects

A current set up in a conductor causes the dissipation of energy in it in the form of heat. A current causes a magnetic field to be set up in the space around it. A current causes chemical dissociation in an electrolyte.

There are numerous other effects that currents and potential differences can produce. Thus, a potential difference applied to a material can cause mechanical stress to be developed in it. A current in a diode can lead to the production of coherent light (laser) in the presence of an appropriate arrangement (an optical resonator, see sections 15.6 and 19.3.7.3). A current through a discharge tube can lead to ionization of the gas in the tube and can produce a glow.

Likewise, currents and voltages in a circuit may be involved in a number of phenomena referred to as *thermoelectric* processes, marked by an interdependence of electrical and thermal effects.

For instance, imagine two wires A and B made of dissimilar metals to be joined to form a loop with two junctions C and D as in fig. 12-49(A). If the junctions C and D are kept at

two different temperatures, then it will be found that a current flows through the circuit, as may be detected by the galvanometer G. This means that, even though no electrical cell is included in the circuit or no varying magnetic field is there, an EMF is generated due to the temperature difference between the two junctions. This is referred to as the thermo-EMF in the circuit, and the production of the thermo-EMF is referred to as the *Seebeck* effect. The direction and magnitude of this EMF depends on the materials of the two wires, as also on the temperatures of the two junctions.

The thermo-EMF developed can be quantitatively expressed in terms of three temperature dependent coefficients, namely the *Peltier* coefficient $\Pi_{AB}(T)$ of the pair of metals A and B, and the *Thompson* coefficients $\sigma_A(T)$ and $\sigma_B(T)$ of the two metals.

A related phenomenon is observed if, in a similar circuit made of the wires A and B, one includes a source of EMF so that a current is sent through the closed loop (fig. 12-49(B)). One then finds that heat is produced in one part of the circuit including one of the junctions C and D, and is absorbed in the remaining part, including the other junction. The effect is seen to be reversed when the direction of the applied EMF is reversed. The same three coefficients as the ones mentioned above determines which of the two junctions gets heated while the other junction gets cooled. This phenomenon of heat being produced and absorbed due an EMF in the circuit is referred to as the *Peltier* effect.

In this book, I will not enter into a more detailed description and analysis of these *thermoelectric* effects, which is found to occur not only in metals, but in *semi-conductors* as well (see chapter 19 for a brief introduction to the physics of semi-conductors). The theoretical explanation of thermoelectric effects, and a quantitative explanation for the thermoelectric coefficients indicated above, is a considerably involved one. The thermoelectric effects occur in *non-equilibrium* situations, analogous to the one where a steady current flows through a circuit, driven by an electrical cell. One has to consider here the *diffusion* of the charge carriers (electrons or holes) due to a concentration gradient of these carriers, as also the *scattering* of these carriers by the vibrating atoms in

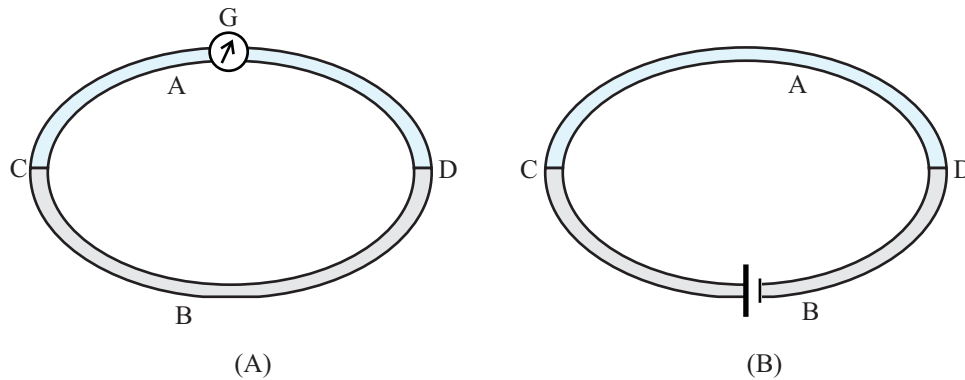


Figure 12-49: Set-up for illustrating thermoelectric effects; (A) Seebeck effect; two dissimilar metals A and B are joined to form a closed circuit, with junctions C and D kept at two different temperatures; this results in an EMF (referred to as a thermo-EMF) being produced in the circuit, causing a deflection in the galvanometer G; (B) Peltier effect; a source of EMF is included in the closed circuit, without C or D being heated or cooled externally; one of the two junctions is then found to get heated and the other cooled.

any and every small region of the circuit. What is more, the scattering by *impurities* and *crystalline imperfections* plays a significant role in the generation of thermoelectric effects.

Thermoelectric effects are put to use in the measurement of temperatures, in temperature controlling devices, and in a number of small power generation and refrigeration units.

Chapter 13

Electricity II: Varying and alternating currents

13.1 Introduction

The electric field was introduced in chapter 11 as being produced by stationary *charges*, while the magnetic field was introduced in chapter 12 as being produced by stationary *currents*. On the face of it, the two appear to be independent entities. A set of stationary charges, for instance, does not produce any magnetic effect and, similarly, a steady current flowing through a wire does not produce any static electrical effect in the form of force exerted on other charged bodies.

In fact, however, there exists a deeper connection between the two. An aspect of this deeper connection is revealed by the fact that the constants ϵ_0 and μ_0 , introduced while describing the forces between stationary charges and those between stationary currents are not independent of one another, but are related as

$$\epsilon_0\mu_0 = \frac{1}{c^2}, \quad (13-1)$$

where c is the velocity of light in vacuum, a fundamental constant of nature. The question may well be asked as to why on earth should an *electrical* constant and a *magnetic*

constant be related to the velocity of *light*, of all things? The answer lies in the fact that considerations relating to stationary electric and magnetic fields are inevitably linked up with those relating to fields *varying with time*.

Consider, for instance, the following situation. A stationary charged particle A produces a force on another stationary charged particle B. Suppose now that the charge A is moved to a new position and held at rest there. In this new position, the force on B exerted by A is different from what it was when A was at its earlier position. In chapter 11, this force was interpreted as an influence set up by A acting on B. The question that comes up then is, how is this influence *transmitted* from A to B? As A is shifted from one position to another, how does B ‘know’ of the change? A similar question can be asked about the influence set up by a steady current, say I_1 , in a conductor resulting in a magnetic force being exerted on another conductor carrying a steady current, say, I_2 . If the former be changed to a new steady value, say, I'_1 , the force on the latter gets changed. But how is this ‘information’ of the current I_1 having been changed passed on to the conductor carrying the current I_2 ?

What has to happen here is that the *fields* themselves have to be responsible in some way in the transmission of the effect caused by the change in the position of a charge or the strength of a current. In other words, the electric and magnetic fields are not just mental constructs for describing the forces between stationary charges and steady currents, since they must themselves mediate in some way in the changes in these forces when the currents and charges are made to change. During these changes in the currents and charges, the fields themselves have to change, and thus the study of steady electric and magnetic fields cannot be separated from that of *time-varying* fields. Indeed, steady electric and magnetic fields are just special instances of time-varying ones. And it is in the context of time-varying fields that one finds that electric and magnetic fields are not as independent of one another as they at first appear to be. One manifestation of this interdependence appears in a set of phenomena termed *electromagnetic induction*, which I will presently turn to.

Indeed, as we will see in chapter 14, electric and magnetic fields varying in time are but

two faces of a *single* dynamical entity - the *electromagnetic* field. The electromagnetic field permeates the whole of space like a great invisible mass of jelly that can quiver and press against objects, push them around, and transfer its own state of agitation to distant places by being set in wavy motions.

Physicists in early days were impressed by these *substance-like* features of the electromagnetic field and imagined an invisible *medium* called ether as the all-pervading substance mediating the interactions between charges and currents. However, while the electromagnetic field is a dynamical system that possesses its own energy and momentum, it does not possess inertial properties like known fluids. The ether concept had eventually to be dropped in favor of a more refined concept of the electromagnetic field, whose wavy motions in vacuum propagate with a velocity that sets the upper limit to all physically realizable velocities in nature.

In other words, steady electric and magnetic fields represent too specialized a set of situations, whereas time-varying fields that come up inevitably when one considers a transition from one stationary situation to another, bring out two completely novel aspects of electric and magnetic phenomena: (a) that electric and magnetic fields are not independent of each other since a time-varying magnetic field generates an electrical effect while, conversely, a time-varying electric field equally well generates a magnetic effect, and (b) that the two make up a single dynamical entity, the electromagnetic field, that carries energy and momentum of its own and can impart part of this energy and momentum to bodies carrying charges and currents, this being precisely the way that a charged particle or a current exerts its influence on a distant charge or current.

13.2 Electromagnetic induction

Fig. 13-1(A) shows a closed loop (C) made up of a conducting wire and a galvanometer (a sensitive current-measuring instrument) placed near a magnet (M). As long as there is no relative motion between the wire loop and the magnet, the galvanometer remains unmoved, not registering a current. If now the magnet is moved toward (or away from)

the loop the galvanometer shows a deflection even though no electrical cell or battery is included in the loop. This shows that an electromotive force is generated in the loop, setting up an electric field along the wire and a current in it just as if an electrical cell were included in the circuit.

A similar effect is produced if, instead of the magnet being moved, the loop is made to move relative to the magnet. Another equivalent way of setting up an EMF in C without an electrical cell being used in the circuit is to make use of a second circuit (fig. 13-1(B)) with an electrical cell included in it, and moving the latter around. What happens here is that the current in the second circuit (D) generates a magnetic field that varies in time as the circuit is moved around. Yet another approach is to set up a *varying* current in the D, in which case the galvanometer in C again registers a deflection.

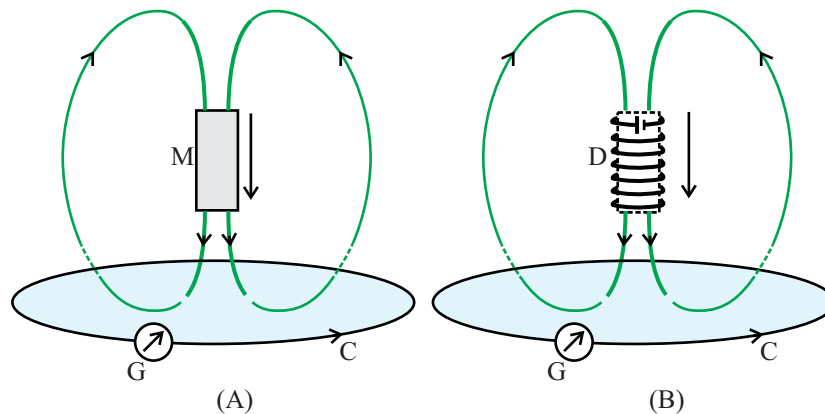


Figure 13-1: Set-up demonstrating electromagnetic induction; (A) a coil C in the field of a magnet M; (B) C in the field produced by a second current-carrying coil D; a few magnetic lines of force are shown schematically in either set-up; as the magnet M or the coil D is made to move relative to C, the changing magnetic field intensity causes an electromotive force to be generated in C, resulting in a deflection in the galvanometer G; the induced EMF is proportional to the rate of change of magnetic flux linked with C.

What is common to all these observations is that a closed loop of wire placed in a magnetic field is endowed with an electromotive force if the magnetic field intensity varies with time. This phenomenon goes by the name of *electromagnetic induction*.

More precisely, it is the change in the magnetic *flux* associated with the coil or a closed circuit that accounts quantitatively for the EMF so produced. This requires a brief

introduction to the concept of magnetic flux - a concept analogous to gravitational or electrical flux (see sections 5.3.1, 11.8.1) which will then lead us to *Faraday's law*.

13.2.1 Magnetic flux

Figure 13-2 shows a surface S in a magnetic field bounded, by a closed curve C . Imagining the surface to be divided into a large number of small area elements, let the area of one such element around the point P on the surface be δs . Let, moreover, the unit vector along the normal to the surface drawn at P be \hat{n} , the normal being chosen in any one of the two possible senses (double-headed arrow in the figure). If the magnetic field intensity at P be \mathbf{B} , then the flux linked with the area element δs is defined to be $\hat{n} \cdot \mathbf{B} \delta s$. The flux linked with the surface S can now be defined as a sum of contributions of all the area elements into which S is imagined to be divided, where all the area elements are assumed to be vanishingly small. The sum then reduces to a *surface integral* of the magnetic field intensity over S , and so the flux is given by the mathematical expression

$$\Phi = \int_{(S)} \hat{n} \cdot \mathbf{B} ds. \quad (13-2)$$

1. The flux as defined above depends on the chosen sense of the unit vectors at the various points of the surface. In the figure, the double-headed arrow indicates the sense chosen. If one chooses the opposite sense for defining the unit vectors, then one will obtain the same result as that given above, but with the opposite sign.

The sense of the normal at various points on the surface S is commonly chosen with reference to a certain *sense of traversal* of the boundary of the surface S , whose choice may depend on the context.

2. An important feature of magnetic fields is that the flux linked with a given surface S depends *only* on the closed boundary C of the surface. In other words, imagining a number of surfaces, all spanning a given closed boundary C , one would obtain the *same* value of flux for all of these.

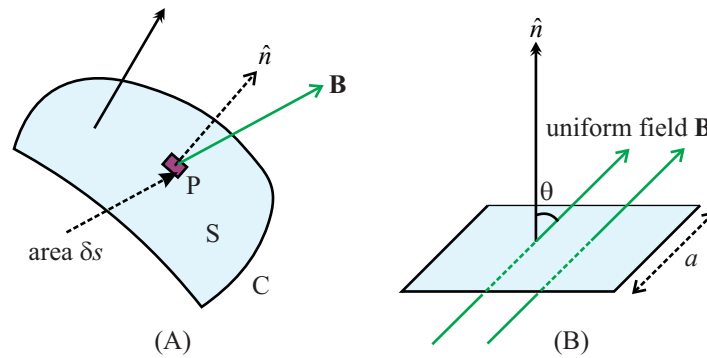


Figure 13-2: (A) Illustrating the definition of magnetic flux linked with a surface S ; C is the closed boundary of the surface; the surface is imagined to be divided into a large number of infinitesimally small area elements; δs is one such element around the point P , at which the magnetic field intensity is \mathbf{B} ; the normal to S at P in the sense of the double-headed arrow is \hat{n} ; the flux (in the sense chosen above) linked with S is then a sum of contributions from all the area elements so obtained, the contribution of the element δs shown being $\hat{n} \cdot \mathbf{B} \delta s$; the sum can be expressed as a surface integral (eq. (13-2)); (B) the particular instance of a closed loop in the form of square of edge length a , placed in a uniform magnetic field of strength B , where the direction of the field makes an angle θ with the normal to the plane of the loop; the flux linked with the loop in the sense of \hat{n} (double headed arrow) is then $|\mathbf{B}| a^2 \cos \theta$.

3. The expression (13-2) tells you why an alternative name for the magnetic field intensity is 'flux density'.

The SI unit of flux is termed the *weber* (W), which is related to the tesla (T) as $W = T \cdot m^2$.

One can now apply the above definition to work out the flux linked with a closed loop of wire or with a closed circuit, since the loop or the circuit traces out a closed curve in space. The flux linked with the circuit is then simply the flux associated with any surface spanning this closed curve. If the loop of wire is made up of a number (say, N) of turns where all the turns are close together, coinciding more or less with some closed curve C , the flux will, to a good degree of approximation, be N times the flux associated with a surface spanning C .

As a particular instance, consider a planar loop of wire placed in a *constant* magnetic field \mathbf{B} (refer to fig. 13-2(B), where a square loop is shown). Then the flux, in the sense of that of the unit vector \hat{n} shown in the figure (associated with an anticlockwise traversal of the boundary of the square), will be $|\mathbf{B}| S \cos \theta$, where S stands for the area of the planar loop, and θ is the angle shown (check this statement out).

Problem 13-1

Consider a closed loop of wire in the form of $N = 100$ turns of a square of edge length $a = 0.05\text{m}$, as in fig. 13-2(B) (only one turn of the loop is shown in the figure), placed in a magnetic field $B = 0.1\text{T}$, where the direction of the field makes an angle $\theta = \frac{\pi}{3}$. Obtain the flux linked with the loop in the sense of \hat{n} (the flux linked in the opposite sense is obtained by adding a negative sign to this).

Answer to Problem 13-1

HINT: The expression for the flux linked is $Na^2B\cos\theta$ which works out to 0.0125W on substituting given values.

13.2.2 Faraday's law of electromagnetic induction

Consider now a closed loop of wire or any closed circuit C placed in a magnetic field, and suppose that the flux linked with the circuit is made to vary with time. This can happen, for instance, in situations indicated in sec. 13.2, as also in various other situations. Let us choose a sense for the unit normal vectors on any surface spanning C , one convenient choice being that corresponding to a chosen sense of traversal of the closed loop under consideration, where the unit normal is assumed to be related to this sense of traversal by the right hand rule (see fig. 13-3). Choosing the sense of the unit normal vectors, and the associated sense of traversal of C , one obtains a definite value for the flux associated with the surface (or, more precisely, with the closed path C).

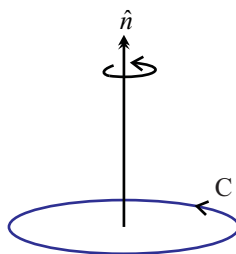


Figure 13-3: A closed path C and a unit normal \hat{n} ; a sense of traversal of C is shown, to which \hat{n} is related by the right hand rule.

Faraday's law of electromagnetic induction then states that the electromotive force \mathcal{E} induced along the closed path C (in the chosen sense of traversal of C) is related to the rate of change of the flux Φ as

$$\mathcal{E} = -\frac{d\Phi}{dt}. \quad (13-3)$$

This law gives us the magnitude of the EMF induced in the circuit under consideration due to the change in the flux linked with it, as also the sense in which this EMF acts. For instance, for a positive rate of change of the flux Φ , the induced EMF acts in a sense opposite to the sense of traversal of the closed path based on which Φ is defined, while for a decreasing Φ , the EMF will be induced in the same sense as that of describing the closed path.

More precisely, choosing a sense of traversal of the closed path C , and the corresponding direction of the unit normal defined at all points on a surface spanning C ,

$$\oint \mathbf{E} \cdot d\mathbf{r} = -\frac{d}{dt} \left(\int \mathbf{B} \cdot \hat{n} \, ds \right), \quad (13-4)$$

where the symbols are self-explanatory by now.

The changing magnetic flux linked with a closed loop of wire then results in an energy being supplied to a charge, say, q if the charge is assumed to make a complete traversal of it, the amount of energy being $q\mathcal{E}$, where it is assumed that the sense in which the charge moves is the same as that in which the EMF is set up. One question that comes up here is, where does this energy come from? A second question is, what happens to this energy?

The answer to the first question is, in brief, *the energy is supplied by whatever causes the flux to change*. And that to the second question is, *this energy goes to heat up the wire and, depending on circumstances, to supply the energy expended in any other form, resulting from the motion of the charge in the circuit*. For instance, it may so happen that the motion of the charge, i.e., the current set up in the wire, results in a mechanical

motion of the latter. Then the energy for this mechanical motion has to come, in the ultimate analysis, from the energy supplied by the agency that causes the flux to change.

13.2.2.1 Lenz's law

Fig. 13-4 shows a closed loop of wire making up a rectangular frame placed in a uniform magnetic field B , where the latter is assumed, for the sake of simplicity, to be in a direction perpendicular to the plane of the wire frame. If now the magnetic intensity is made to increase with its direction being kept unaltered, the flux Φ (in the sense of B) linked with the wire will increase with time and an EMF will be induced in the wire loop in a sense opposite to that related to the sense of B according to the right hand rule. A current will then flow through the wire in the direction indicated by arrows in the figure.

A curious thing can now be noted. The induced current flowing through the wire loop will set up its own magnetic field. Significantly, this field will be directed *oppositely* to the field we started with, along the dotted arrow in the figure. The *flux* Φ' due to this field, in the sense of the *initial* field, will then be negative, i.e., in other words, it will tend to cancel the increase in flux Φ that caused the induced current in the loop to appear in the first place.

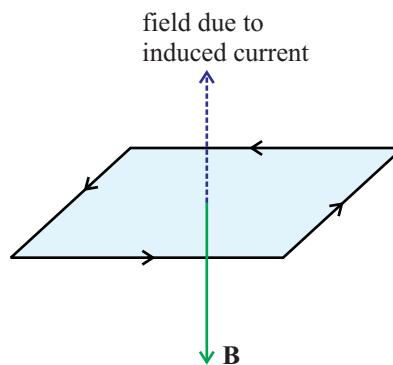


Figure 13-4: Illustrating Lenz's law; as the strength of the field B is made to increase, the induced EMF in the closed loop of wire sets up a current in the direction of the arrows which, in turn, results in a magnetic field in the direction of the dotted arrow; the flux due to this field tends to cancel the increase in flux due to B that causes the induced EMF in the first place.

Here, then, is an interesting relation between *cause* and *effect* (or the *response* to the

cause). The increase in the flux Φ due to the field B is here the cause setting up the induced EMF in the coil. The resulting current and the flux Φ' due to *this* current may then be looked upon as the effect, or the response to the cause. The summary is, therefore, as follows: *the effect tends to annul the cause*.

Electromagnetic induction may result in other modes of response as well. Another instance of response will be found in section 13.2.4 where it will be seen that the response consists of a *rotation* of the wire frame provided the latter is free to turn about an axis. Once again, the sense of rotation will be such that the change in flux causing the response tends to get annulled.

This relation between cause and effect, inherent in electromagnetic induction and expressed by the negative sign associated with the expression for the induced EMF in eq. (13-3), is referred to as *Lenz's law*. While being a useful principle in practical considerations, it is not a fundamental principle in itself, being a consequence following from Faraday's law.

13.2.2.2 Motional EMF

Fig. 13-5(A) shows a conducting rod AB moving in a magnetic field, where the direction of motion of the rod is assumed for the sake of simplicity to be perpendicular to its length (in the plane of the diagram) and the direction of the magnetic field is assumed to be perpendicular to both the length and the direction of motion of the rod (pointing, say, *into* the plane of the diagram). The field is, moreover, assumed to be uniform, again for the sake of simplicity.

What happens in this case is that the motion of the rod in the magnetic field causes an *electric field* to appear along the length of the rod. One way to see why this has to be so is to imagine a carrier (or a number of carriers) of charge $-q$ ($q > 0$) in the rod. This charge shares the velocity (say, v) of the rod and experiences a Lorentz force (see sec 12.8.4) $-q\mathbf{v} \times \mathbf{B}$ which, under our present assumptions, is of magnitude qvB and acts along the length of the rod in the direction from A to B in the figure. However, the rod not being a part of a complete circuit, no *current* can flow through it, and hence this

force has to be annulled by an equal and opposite force appearing along the rod. This force is brought into play by the electric field appearing along the length of the rod. The mechanism responsible for the setting up of the field is the accumulation of charge at the two ends of the rod (due to the operation of the Lorentz force), which continues so long as the Lorentz force is not balanced by the force due to the electric field.

If the electrical intensity along the length of the rod be denoted by \mathbf{E} , then the condition for the balancing of the two forces gives

$$\mathbf{E} = -\mathbf{v} \times \mathbf{B}, \quad (13-5)$$

which, under our present assumptions, corresponds to a potential difference

$$V_A - V_B = vBl, \quad (13-6)$$

between the ends of the rod, with the end A in the figure being at a higher potential compared to the end B.

Problem 13-2

Check eq. (13-6) out.

Answer to Problem 13-2

HINT: Refer to eq. (11-7) with \mathbf{r}_0 and \mathbf{r} corresponding to the two ends of the rod. For a uniform electric field \mathbf{E} the potential difference between two points, say, A and B, is obtained from this formula as, $V_A - V_B = -\mathbf{E} \cdot \mathbf{l}$, where \mathbf{l} is the vector directed from B to A. Using eq. (13-5), this gives (13-6).

Fig. 13-5(B) depicts a similar situation but now with the rod AB being made to slide along two arms of a pair of rails, closing an electrical circuit. In this case, the motion of the rod results in an *electromotive force* \mathcal{E} being set up in the circuit *in the sense shown by the arrows*, where \mathcal{E} , termed the *motional EMF*, is given by the same expression as in

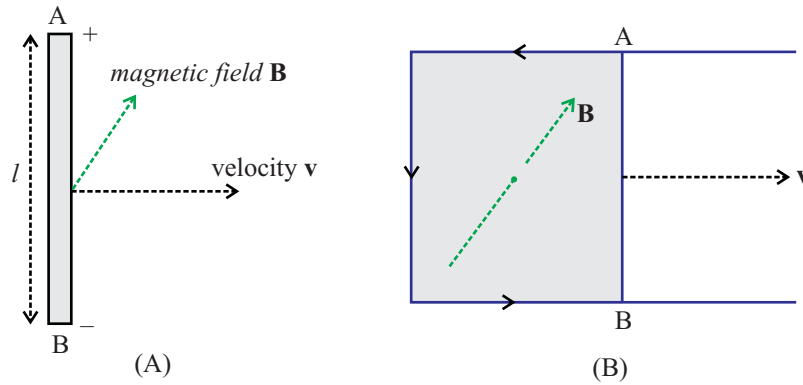


Figure 13-5: Illustrating the idea of motional EMF; (A) a conducting rod moving in a magnetic field; a potential difference appears between the ends A and B of the rod so that no current can flow through it; under simplifying assumptions, this potential difference is given by eq. (13-6); (B) the rod slides along two arms of a pair of rails, closing an electrical circuit; an electromotive force now appears in the circuit, given by (13-7).

eq. (13-6):

$$\mathcal{E} = vBl. \quad (13-7)$$

This is consistent with Faraday's law (13-3) since the flux linked with the circuit is BA , where A denotes the area enclosed by the rod and the rails, and this increases at the rate Blv (check this out; observe that the sense of the induced EMF is related oppositely to the direction of B as compared to the one implied by the right hand rule since here the flux in the sense of B is made to increase with time).

As a result of this induced EMF operating in the circuit, a current (I) flows in it, and a consideration of the energy accounting in the circuit (where the rate of supply of energy by the source of EMF equals the total rate of energy dissipation (see section 12.5.3)), gives

$$\mathcal{E} = vBl = I(R + r). \quad (13-8)$$

Here r stands for the resistance of the rod while R denotes the resistance of the rest of the circuit.

The supply of energy balancing the dissipation in the circuit comes, in the ultimate analysis, from the work done by the force required to make the rod slide on the rails because it is this force that is responsible for the generation of the EMF. The force on the rod, carrying a current I , due to the magnetic field of strength B is (see sec. 12.8.3) IlB in a direction toward the left in fig. 13-5(B), and an equal and opposite force is to be exerted by an external agency so as to make the rod slide in the direction shown in the figure. The rate of work done by this force will then be $vBlI$, which is precisely the rate of energy dissipation $I^2(R + r)$.

The potential difference between the ends of the rod ($V_A - V_B$) is now no longer given by eq. (13-6), since it now carries the current I (recall how the potential difference between the terminals of an electrical cell gets diminished from the open circuit voltage of the cell due to the internal drop of voltage) and one has, instead,

$$V_A - V_B = vBl - Ir = \mathcal{E} - Ir, \quad (13-9)$$

which is obtained from eq. (13-8) by noting that, according to Ohm's law, applied to the rest of the circuit, $V_A - V_B = IR$. A simple-minded application of Ohm's law, applied to the *rod* would give $V_A - V_B = -Ir$, which would be a contradiction. Recall, however, that Ohm's law for any given part of the circuit is an expression for the energy balance principle considered for that part. For the rod, as opposed to the rest of the circuit, one has to consider the magnetic force on the carriers due to the motion of the rod and the work done by the external agency (the one that causes the motion of the rod), in order to arrive at the energy balance equation.

1. The carriers possess two kinds of motion - a motion they share with the rod as a whole, and the drift motion along the rod. It is the former that is to be taken into account while obtaining the potential difference between the ends of the rod. In the frame of reference of the rod, however, one need not consider the magnetic force on the carriers and the potential difference between the ends B and A of the rod will be Ir .
2. The statement that electric and magnetic fields are not independent concepts, can

be interpreted in the following manner. The distinction between electric and magnetic fields is *frame-dependent*: what appears to be a magnetic field in one frame appears in another frame to be an electric field, or, more generally, a combination of an electric and a magnetic fields. Thus, the description of the fields in the frame of reference of the rod will differ from the way the same fields are described in the frame of reference of the rest of the circuit.

13.2.3 The principle of DC and AC generators

The set-up of fig. 13-5(B) illustrates, in principle, the basic concept underlying an electrical *generator*, the latter being a device where a mechanical motion is imparted to one part (the rod in the instance of fig. 13-5(B)) of a closed circuit, generating an EMF (\mathcal{E}) in it, where the energy necessary to impart the motion is responsible in maintaining the EMF in the circuit. If the circuit be an open one, i.e., the terminals of the moving part are not joined by an external conducting path, then the potential difference between these terminals equals the EMF that would be produced if the connection through the external path were made (equations (13-6) and (13-7)).

Recall that this is true of the EMF (\mathcal{E}) produced by an electrical cell as well. The open circuit potential difference between the terminals of the cell equals \mathcal{E} , but if the terminals of the cell are connected through a resistance R , then the potential difference between the terminals gets reduced to $\mathcal{E} - Ir$, where I is the current set up in the closed circuit and r is the internal resistance of the cell. In the set-up of fig. 13-5, the moving rod can be looked upon as the source of the EMF, analogous to the Galvanic cell, and its resistance r is analogous to the internal resistance of the cell. As the device supplies a current I through the circuit, the potential difference between the ends of the rod gets reduced (eq. 13-9) by Ir .

In the set-up shown in fig. 13-5, if the rod be made to slide with a uniform velocity in a uniform magnetic field, then the EMF in the circuit will also remain constant and unidirectional. This corresponds to the moving rod operating as a source of *DC* EMF (a *DC source* in brief). If on the other hand, the rod were made to undergo an oscillatory

motion on the rails then the EMF would not remain constant. By properly controlling the oscillatory motion, the EMF in the circuit may be made to vary sinusoidally, sequentially undergoing a change in direction. This corresponds to the rod operating as a source of AC EMF or what is commonly referred to as an *AC source*.

In practice, however, this is not how a DC or an AC generator is made up. Instead of one single rod, a number of rods are used, and instead of a translational motion in a magnetic field, the rods are made to *rotate* about an axis in the field. Still, the basic principle remains the same, namely, the generation of an EMF by means of *electromagnetic induction*.

13.2.3.1 Conducting frame rotating in a magnetic field

Figure 13-6(A) shows a frame made up of two conducting rods AB and CD, the terminals (T_1 , T_2) at the two ends A and D being open. The frame is situated in a uniform magnetic field of strength B (dotted line with arrow). Suppose that the frame is made to rotate about the axis XY with an angular velocity ω , where, at time $t = 0$, the plane of the frame is perpendicular to the direction of the field.

Even though there is no external connection to the terminals (T_1 and T_2) of the frame, for the purpose of analysis it is convenient to imagine that an infinitely large resistance is connected between the terminals so as to make up a complete circuit, of which only the part made up of the frame lies in the region occupied by the magnetic field. The flux linked with the circuit at time t is then $AB \cos \omega t$ (check this out; the sense of traversal of the circuit based on which the flux is defined, is to be taken as that from D to A through C and B), where A stands for the area of the frame. Since the flux changes with time, an EMF is induced in the circuit, the EMF at time t (in the sense of traversal of the circuit mentioned above) being

$$\mathcal{E} = AB\omega \sin \omega t, \quad (13-10)$$

(check this out).

Suppose now that an external connection between the terminals T_1 and T_2 is established as in fig. 13-6(B) by means of the ring-like conducting contacts R_1 and R_2 , the resistance in the external circuit being R . The terminals of the external circuit are connected to the rings by means of two conducting brushes (tiny rectangles in the figure) that press against the respective rings, where the latter are fixed rigidly with the rods so that the ring R_1 is in permanent electrical contact with BA and R_2 with CD. If the resistance of the frame itself be r , then the current through R from the end E to F at time t will be

$$I = \frac{AB\omega}{R+r} \sin \omega t, \quad (13-11)$$

which varies sinusoidally and sequentially changes direction. Such a current is referred to as an *alternating current* (AC in short) with angular frequency ω and *amplitude* $I_0 = \frac{AB\omega}{R+r}$. The rotating frame here acts as an AC source, commonly referred to as an AC generator or *alternator*.

In practice, alternators of large power generating capacity are often built with the magnetic winding (mounted on a core with appropriately cut pole-pieces) constituting the rotating member ('rotor') and the armature coil (made of conductors across which the AC voltage is produced) constituting the stationary member ('stator'). Such an assembly has a number of notable advantages over the stationary-field-rotating armature type assembly since the the external connection carrying the load can be made to remain in fixed contact with the armature coil while the contact with the rotor coil, made up of slip rings and brushes, has to handle a lower power supply. The theory of power generation with such an assembly remains the same, being based essentially on the relative motion between the field and the armature.

A second mode of connection of the external circuit is shown in fig. 13-6(C). Here a pair of conducting segments C_1 and C_2 , insulated from each other, are made use of, to which the two ends E and F of the external circuit are brought in contact by means of conducting brushes. As the frame ABCD rotates in the magnetic field, along with the segments C_1 and C_2 (termed the *commutator* segments) rigidly connected in contact with

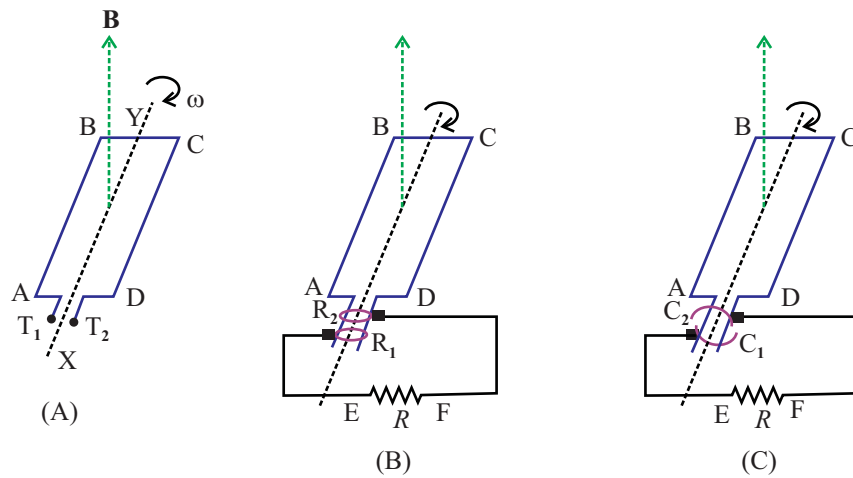


Figure 13-6: A conducting frame, made up of conductors AB and CD, rotating about the axis XY in a uniform magnetic field B ; (A) the terminals T_1 and T_2 are open, which is equivalent to there being a connection between the terminals through an infinitely large resistance; the flux linkage with the circuit changes sinusoidally, and the frame acts like a source of AC voltage; (B) connection is established between the terminals of the frame by means of the rings R_1 and R_2 through a resistance R , in which an alternating current is set up; (C) connection is established by means of the commutator segments C_1 and C_2 , which makes the current in R unidirectional, from E to F; in (B) and (C), electrical contact between the terminals of the frame and those of the external circuit is established by means of brushes shown in the figure as tiny rectangles.

the arms AB and CD respectively, its terminals T_1 and T_2 (one at the end of the rod AB, and the other at the end of CD, see fig. 13-6(A)) are alternately brought into electrical contact with the terminals E and F of the external circuit.

Suppose that at some instant of time the induced EMF in the frame acts from D toward A and that, at this instant, T_1 is in contact with E through the commutator segment C_1 by means of the brush with which E is connected. The other terminal T_2 , at the end of the arm CD, will then be in contact with F by virtue of C_2 being momentarily in contact with the other brush. The current in the external circuit will then be from E to F. At a later instant, as the arms AB and CD of the frame interchange positions, T_1 comes in electrical contact with F since C_1 (which rotates in rigid electrical contact with the arm AB) will now press against the brush connected to F and T_2 will similarly be in electrical contact with E while, at the same time, the induced EMF will now act from A to D. The current in the external circuit will thus once again be from E to F. In other words, this mode of connection results in a unidirectional current being set up in the external

circuit, which is how a DC generator acts.

Though unidirectional, the current in the external circuit in fig. 13-6(C) rises and falls owing to the fact that the rate of change of flux linked with the frame is not uniform. It is minimum, for instance, when the plane of the frame is perpendicular to the direction of the magnetic field (for instance, at $t = 0$ in eq. (13-11); because of the connection through the commutator segments, the current in the direction from E to F cannot be negative, and is given by the magnitude of the expression in the right hand side) and is maximum when the plane of the frame makes an angle $\frac{\pi}{2}$ with its position shown in fig. 13-6(C) (for instance, at $t = \frac{\pi}{2\omega}$).

In order to make the current in the external circuit more uniform, a number of frames like the one shown in fig. 13-6(C) are used, all connected rigidly with one another and rotating coaxially in the magnetic field. The electrical connections between the terminals of the frames, the commutator segments, and the brushes being made in accordance with a number of practical considerations.

For our purpose, the basic idea can be explained by means of a simplified connection scheme. Instead of talking in terms of a number of frames, it will be more convenient for the present to talk in terms of a number of pairs of rods, like AB and CD in the figure 13-6. One end of each rod belonging to a pair (like T_1 for AB and T_2 for CD) is connected in permanent electrical contact to a commutator segment (while one end of the other rod of the pair is similarly connected to another commutator segment), where we assume for the sake of simplicity that there are just two such segments attached rigidly with the entire assembly of rods. The remaining ends of the pair of rods under consideration are connected electrically to each other.

This means that the frame made up of each pair has its terminals (like T_1 and T_2 above) connected between the commutators C_1 and C_2 , and thus all these frames, considered as so many conductors, are in parallel connection. If contact is established with the terminals of the external circuit by means of brushes as above and the relative orientations of the frames are appropriately fixed, then the entire rotating assembly acts as a unidi-

rectional source of EMF, where the fluctuations due to the individual frames get evened out (refer to note in problem 13-3). This constitutes the operating principle of the DC generator. In reality, the rotor assembly (a rigid laminated core with the rods mounted in parallel slots) includes as many commutator segments as the number of rods, with one end of each rod in permanent contact with one segment, and the segments come in contact with the fixed brushes in rapid succession, thereby producing the desired effect of rotating coils mounted in parallel where a constant DC voltage is produced across the brushes and the effective internal resistance of the generator is lowered.

An AC current or voltage is one that varies sinusoidally with some given frequency. Basic concepts relating to such currents and voltages, and to the analysis of circuits carrying AC currents will be found in later sections of this chapter.

Problem 13-3

Consider two electrical cells, of EMF's E_1 , E_2 and internal resistances r_1 , r_2 , connected in parallel across an external resistance R , as in fig. 13-7. Find the current through the external resistance, as also that through each of the two cells. Interpret the result by referring to the special case when the two cells are of identical internal resistance r , and generalize to the case of N such cells ($N = 2, 3, \dots$), all of identical internal resistance, connected in parallel.

Answer to Problem 13-3

HINT: Considering the mesh made up of the cell of EMF E_1 and internal resistance r_1 , and the external resistance R , and invoking Kirchhoff's second principle to this, one obtains (refer to fig. 13-7) $IR + I_1 r_1 = E_1$, i.e., $I_1(R + r_1) + I_2 R = E_1$ (check this out) while, considering the mesh made up of the other cell and the external resistance R , one obtains $I_1 R + I_2(R + r_2) = E_2$. These two relations yield the solution $I_1 = \frac{E_1(R+r_2)-E_2 R}{R(r_1+r_2)+r_1 r_2}$, $I_2 = \frac{E_2(R+r_1)-E_1 R}{R(r_1+r_2)+r_1 r_2}$, and hence the main current $I = \frac{E_1 r_2 + E_2 r_1}{R(r_1+r_2)+r_1 r_2}$. In the special case when the two cells are of identical internal resistance r , the expression for main current simplifies to $I = \frac{E_1 + E_2}{2R + r}$, which can be interpreted as the current sent out by an equivalent electrical cell of EMF $\frac{E_1 + E_2}{2}$ and internal resistance $\frac{r}{2}$, the latter being the equivalent resistance of two resistances, r each, connected in parallel. One can generalize this by saying that, In the case of N number of cells of EMF's E_1, E_2, \dots, E_N , all of identical

internal resistance r , the combination acts as a single cell of EMF $E = \frac{1}{N} \sum_{k=1}^N E_k$ and of internal resistance $\frac{r}{N}$.

NOTE: This is of relevance in the connection of the pairs of rods forming the frame of conductors in a rotating magnetic field discussed above in connection with the DC generator, where each pair of rods instantaneously acts as a source of EMF, all the different sources formed by the various pairs being in parallel; the equivalent EMF, being the average of the EMF's due to the various different pairs, remains constant in time, and the equivalent internal resistance is reduced considerably.

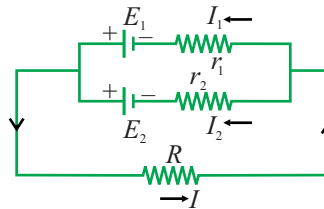


Figure 13-7: Two electrical cells, of EMF's E_1 , E_2 and internal resistances r_1 , r_2 , connected in parallel across an external resistance R ; currents I_1 , I_2 flow through the two cells, while the main current flowing through R is $I = I_1 + I_2$ (Kirchhoff's first principle); expressions for these can be worked out as in problem 13-3; in the case of N number of cells, of EMF's E_1, E_2, \dots, E_N , all of identical internal resistance r , the combination acts as a single cell of EMF $E = \frac{1}{N} \sum_{k=1}^N E_k$ and of internal resistance $\frac{r}{N}$.

13.2.4 Rotating magnetic field: AC motors

Fig. 13-8 shows a frame made of conducting rods placed in a magnetic field, with the latter being a *rotating* one. At any given instant of time, say $t = 0$, the magnetic field is uniform, with the lines of force along the double-headed arrow. At a later time t , the field lines are along the dotted arrow, having been rotated from their former position by an angle ωt about the axis XY, corresponding to an angular velocity ω about XY. All the while, the magnitude of the magnetic field intensity remains constant, its direction getting rotated at a constant rate.

Such a rotating magnetic field is made use of in AC *motors* where electrical energy fed to a system of coils (or *windings*, as these are often termed) results in a rotational motion in a second body about a fixed axis. This rotational motion is then made use of in running

machines of various descriptions for practical purposes. In fig. 13-8 the frame made up of the conducting rods is assumed to be the rotating body for the sake of simplicity. A current from an external DC source may be fed to the frame, as in a *synchronous motor*. In an *asynchronous motor* no current is fed from an external source, while the rotating magnetic field induces an EMF in the frame.

A rotating magnetic field can be set up with the help of three sets of coils, each being fed with an alternating current from an AC generator so that each coil produces its own magnetic field that varies sinusoidally. The resultant field is then the vector sum of the three fields so produced. The set-up is designed in a clever way so that the three magnetic fields are of the following forms

$$\begin{aligned} \mathbf{B}_1 &= B_0 \hat{i} \cos(\omega t), \\ \mathbf{B}_2 &= B_0 \left(-\frac{1}{2} \hat{i} + \frac{\sqrt{3}}{2} \hat{j} \right) \cos\left(\omega t + \frac{2\pi}{3}\right), \\ \mathbf{B}_3 &= B_0 \left(-\frac{1}{2} \hat{i} - \frac{\sqrt{3}}{2} \hat{j} \right) \cos\left(\omega t - \frac{2\pi}{3}\right), \end{aligned} \quad (13-12)$$

where \hat{i} and \hat{j} stand for the unit vectors along the x- and y-axes respectively of an appropriately chosen Cartesian co-ordinate system..

Working out the vector sum of these, the resultant field is seen to be given by

$$\mathbf{B} = \frac{3B_0}{2} (\hat{i} \cos(\omega t) - \hat{j} \sin(\omega t)), \quad (13-13)$$

which has a constant magnitude ($B = \frac{3}{2}B_0$), and rotates with angular velocity ω about the z-axis in a left handed sense (check this out; comparing with fig. 13-8, the unit vector \hat{k} along the z-direction is directed from Y to X). The three coils producing the fields as in (13-12) have to have their planes inclined at an angle $\frac{2\pi}{3}$ with one another, and the *phases* of the AC currents fed to successive coils are to differ from one another by $\frac{2\pi}{3}$ as well (see sec. 13.5.1.1 for an introduction to the concept of *phase* of an AC current or voltage).

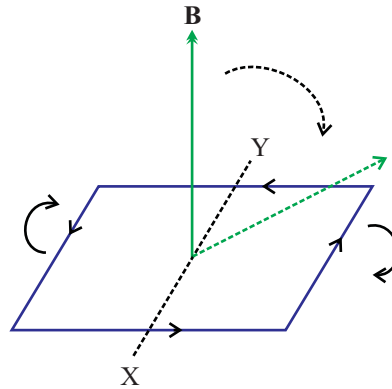


Figure 13-8: A frame made up of conducting rods placed in a rotating magnetic field; the direction of B is made to rotate about the axis XY with an angular velocity ω , causing a change in the flux linked with the frame; as a result, a current will be induced in the frame and forces will be exerted on the arms of the frame by virtue of this current in the conductors being in the field B ; if the frame be free to rotate about the axis XY , then it will rotate in the same sense as B so as to catch up with it, in accordance with Lenz's law.

Problem 13-4

Establish the formula (13-13) by making use of the relations (13-12). Specify the co-ordinate axes by referring to fig. 13-8.

Answer to Problem 13-4

HINT: Referring to fig. 13-8, we choose the x - and y -axes of our co-ordinate system in a plane perpendicular to the line XY , about which the rotation of the magnetic field is desired to take place, the latter line being then chosen as the z -axis pointing in the direction from Y to X . The x -axis is chosen in a direction perpendicular to the plane of the frame, pointing upward, while the y -axis is chosen in the plane of the frame, perpendicular to XY , pointing toward the left. The x -component of the vector sum of the fields in relations (13-12) is given by $B_x(t) = B_0[\cos \omega t - \frac{1}{2}(\cos(\omega t + \frac{2\pi}{3}) + \cos(\omega t - \frac{2\pi}{3}))] = \frac{3}{2}B_0 \cos \omega t$ while, similarly, the y -component is seen to be $B_y(t) = -\frac{3}{2}B_0 \sin \omega t$, thereby verifying formula (13-13).

13.2.4.1 The synchronous motor

In a synchronous AC motor, a DC current is fed to the conductors making up the frame in a direction indicated by the arrows in fig. 13-8, while the rotating magnetic

field is produced by three sets of coils fed with AC currents. The current loop in the frame is equivalent to a magnetic dipole, where the dipole moment is along a direction perpendicular to the plane of the frame, related to the direction of the current in the right hand sense. As can be seen from the formula (12-74b) in section 12.8.12.3, the effect of a magnetic field on a dipole is to exert a torque on the dipole so that the dipole tends to get aligned with the magnetic field. In the present context, the torque will make the frame rotate about the axis XY so that the perpendicular to the plane of the frame tends to rotate in step with the magnetic field.

13.2.4.2 The asynchronous motor

In an asynchronous AC motor the supply of a current from an external source to the frame is not necessary because what happens here is that the rotating magnetic field *induces* an EMF in the closed circuit formed by the frame, resulting in an induced current in the direction shown in fig. 13-8.

If the frame were stationary in the position shown in fig. 13-8, then the flux linked with the frame would be given by the expression $AB \cos \omega t$ at time t , where A stands for the area of the frame and B denotes the magnitude of the magnetic field intensity, the latter being a constant ($= \frac{3B_0}{2}$; refer to eq. (13-13)). The flux linked with the frame would thus change sinusoidally from AB to $-AB$ and back, inducing the EMF in the frame, and causing the frame to rotate in step with the field so as to annul the flux change. In reality, the frame rotates at a rate somewhat slow compared to the rotation of the field (see below) and the flux changes at a slow rate.

The magnetic dipole moment due to this current circulating in the frame then tends to get itself aligned with the rotating magnetic field due to the torque exerted by the latter. It is this resulting rotation of the frame that makes the device operate as a motor.

One crucial feature underlying the operating principle of this machine is that the dipole moment is generated by the rotating magnetic field itself, as a result of which the rotation of the frame cannot be exactly synchronized with that of the magnetic field since, for

an exactly synchronized rotation, there would be no electromagnetic induction possible (the flux linked with the frame would always be zero) and hence no torque. The rotation of the frame would then be useless from a practical point of view since, in order to act as a motor, it would have to move other bodies (such as machine parts), overcoming various resistive effects, the latter exerting a torque on the rotating frame tending to slow it down. In order that the rotating field can exert a forward torque on the frame overcoming this retarding torque, the frame has to lag behind the rotation of the field to some extent.

This lag between the rotation of the field and that of the frame accounts for the phrase 'asynchronous'. In the synchronous machine, on the other hand, the rotation of the frame occurs in step with that of the field.

13.2.5 The principle of DC motors. Back EMF.

In principle, the DC motor is simply a DC generator run in the reverse. If you send a direct current through a frame made of conducting wire or rods in a magnetic field, the forces exerted by the field on the rods making up the frame result in a torque acting on the frame, making the latter rotate about an axis. This rotational motion of the frame can then be made use of in imparting rotational or translational motion to various bodies of interest, supplying energy to these in the form of work.

Fig. 13-9 depicts a frame made of conducting rods in a magnetic field (in the direction of the dotted arrow), where the frame is capable of rotating about the axis XY. If a current is made to flow through the frame in the direction shown by the arrows on the four arms of the frame, then the latter can be looked upon as a magnetic dipole oriented along the double-headed arrow. As indicated in chapter 12, the magnetic field exerts a torque on the dipole, tending to align the latter along the direction of the field. This makes the frame rotate about the axis XY in the sense shown by the semi-circular arrow. It is this rotation of the frame that is achieved in a DC motor.

Referring now to fig. 13-6(C), imagine that a source of EMF is included in the external

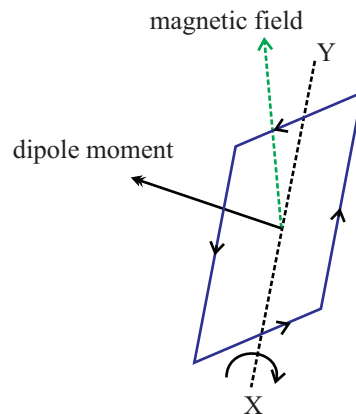


Figure 13-9: Illustrating the torque on a frame made up of conducting rods, where the frame is placed in a magnetic field, and a current is made to flow through the rods in the direction of the arrows; this results in the frame behaving as a magnetic dipole with the dipole moment related to the direction of flow of current in a right-handed sense, and a torque on the frame, causing it to rotate about the axis XY in the direction shown by the semi-circular arrow.

circuit (no such source is actually shown in the figure since it is meant to illustrate the principle of a DC generator) so as to send a current through the conducting frame. As shown in the figure, the plane of the frame is perpendicular to the direction of the magnetic field at the instant under consideration. This, however, is an exceptional position for the frame, because the disposition of the commutator segments and the brushes is so arranged (not shown in the figure) that both the commutator segments touch each brush at this particular position, and the frame itself is cut off from the external DC supply.

This is made necessary by the fact that the back EMF (see below) in the frame is zero at this position and, unless the frame were cut off from the external supply, it would be damaged by the large current that the latter would send through it.

The frame crosses this special ‘neutral’ position by virtue of previously acquired rotational motion, which makes it continue to rotate due to inertia. Let us then imagine a position of the frame slightly rotated from the one shown in the figure (fig. 13-6(C)) in the sense of the bent arrow, when the disposition of the commutators with respect to the brushes is as shown (each commutator in contact with one single brush now).

Suppose that the external DC source is so connected as to send a current through the frame directed from A towards D, resulting in a magnetic dipole aligned oppositely to the magnetic field. This means that, at the instant under consideration, there is a small torque on the frame in the direction of the bent arrow, making the frame continue in its rotational motion. the magnitude of the torque increases as the rotation continues and becomes maximum after the completion of a quarter turn, when the plane of the frame becomes parallel to the magnetic lines of force. Thereafter, the torque decreases, but the rotation continues.

After the frame makes a half turn from the position we started from and crosses the next neutral position (again by virtue of its acquired rotational motion), the electrical contacts of the arms AB and CD with the terminals of the external source (E and F in the figure where the external DC source is not shown) get reversed. As a result, the direction of the magnetic dipole *in space* remains the same as before (with reference to the rotating frame, on the other hand, the direction of the dipole flips over) and hence the torque continues to act in the same sense as before. The frame continues to rotate with a certain angular speed determined by the resistive torque faced in the course of rotation (in reality, for a single frame formed of a pair of rods as shown in the figure, the angular speed is variable), and the basic motor action is achieved.

For the frame shown in figure 13-6(C), the magnitude of the torque evidently fluctuates between zero and a maximum value (μB , where μ is the magnitude of the magnetic dipole moment, and B stands for the magnitude of the magnetic field intensity), which is not appropriate for a motor. This is remedied by making use of a number of frames, each of the type shown in the figure, connected to the commutator segments in parallel and all together making up a single rigid framework, in much the same way as in a DC generator. At any given instant of time, the various different frames experience torques of varying magnitudes, but the resulting torque on the rigid rotating framework remains almost constant in time.

The continued rotation of the rigid framework requires an expenditure of energy for overcoming various kinds of mechanical resistances to the rotation, such as the ones

caused by frictional forces and torques on it and on systems mechanically coupled to it. The supply of this energy comes, in the ultimate analysis, from the source of external EMF supplying the current to the conducting rods making up the framework. A steady state ensues as the framework picks up a speed of rotation that ensures equality of the power supplied by the source of EMF and the power losses due to heat dissipation in the resistive elements and various frictional losses.

13.2.5.1 DC motor: back EMF

As the frame made up of the conducting rods keeps on rotating in the magnetic field, the change in the flux linked with it causes an induced EMF to appear in it. This is referred to as the *back EMF* in the motor. According to Lenz's law, the direction of operation of this back EMF will be such as to oppose the rotation of the frame. The external source of EMF has to overcome this back EMF so as to drive the current through the frame necessary to maintain the torque to keep the frame rotating.

If \mathcal{E} denotes the EMF of the external source and r the total resistance in the external circuit including the internal resistance of the source, then the potential difference between the two terminals of the frame is given by

$$\mathcal{E} - Ir = V, \quad (13-14a)$$

where I is the current flowing through the circuit. This potential (V) between the terminals of the rotating frame has to be larger than the back EMF \mathcal{V} so as to make the current flow through it, and one can write

$$V - \mathcal{V} = IR, \quad (13-14b)$$

where R stands for the resistance of the frame. One therefore has

$$\mathcal{E} - \mathcal{V} = I(r + R), \quad (13-14c)$$

and, as a corollary,

$$\mathcal{E}I = \mathcal{V}I + I^2R + I^2r. \quad (13-15)$$

This last equation is important from the energy point of view. It tells you that the rate at which the external source of EMF supplies energy to the entire set up can be broken up into three parts. Two of these correspond to the rate of generation of heat (I^2r) in the external circuit and that (I^2R) in the resistance of the rotating frame. The principle of conservation of energy then implies that the third part ($\mathcal{V}I$) has to be the energy required to overcome the mechanical resistance to the rotation of the frame, which includes the mechanical energy imparted to other bodies that may be coupled to the rotating frame. One finds that this energy necessary to maintain the rotation of the frame is simply the back EMF times the current flowing through it.

Problem 13-5

Fig. 13-10 shows a set-up with a straight conducting wire (call it W) lying between two fixed conducting rails and a battery of EMF E . The wire can slide over the rails set at a distance l apart, and the entire set-up lies in a magnetic field, with uniform intensity B perpendicular to the plane of the rails (i.e., the plane of the figure; the magnetic field lines point into this plane). Assuming the force of limiting friction between the wire and the rails to be F_0 , find the current in the wire, and set up the energy balance equation for the system .

Answer to Problem 13-5

HINT: If the current in the wire be I , the magnetic field exerts a force $F = BIl$ on it in the direction shown with the double-headed arrow. If v be the velocity of the wire in this direction then the back EMF in the circuit due to the changing flux through the closed loop made up of the rails and the wire, is Bvl (reason this out; this is the same as the motional EMF we came across for the set-up of fig. 13-5). Thus, $E - Bvl = Ir$, where r stands for the total resistance in the loop. In the steady state, as the wire slides along the rails with a constant velocity and carries a steady current, the net force on the wire is zero, i.e., $F = F_f$, the force of friction impeding the motion of the wire on the rails. One now has to distinguish between two cases.

(A) If $F_0 > \frac{BlE}{r}$, the force of friction $F_f = BIl$ is less than the limiting value and the wire remains stationary on the rails, i.e., $v = 0$, and hence $I = \frac{E}{r}$. The back EMF is actually zero, and the energy balance equation reads $P = EI = I^2 r$, i.e., the power delivered by the source is used up as energy dissipated in the circuit.

(B) On the other hand, if $F_0 < \frac{BlE}{r}$, the force of friction attains the limiting value, i.e., $BIl = F_0$. In other words, one has, $I = \frac{F_0}{Bl}$, and the wire moves with a velocity v given by $v = \frac{E - Ir}{Bl} = \frac{E - \frac{F_0 r}{Bl}}{Bl}$. The energy balance equation reads $P = EI = BvlI + I^2 r = F_0 v + I^2 r$, which means that the power supplied by the source goes in part to overcoming the energy loss due to friction (this part being accounted for by the back EMF in the circuit), the rest being dissipated as $I^2 r$ loss.

The principles underlying the operation of this set-up are analogous to those in a DC motor. The case (A) then corresponds to the motor refusing to run because of overload. The case (B) corresponds to the motor running with a steady angular velocity, with the external source of EMF delivering energy to the moving parts (by way of overcoming the resistive forces) and, at the same time, providing for the energy dissipated in the circuit.

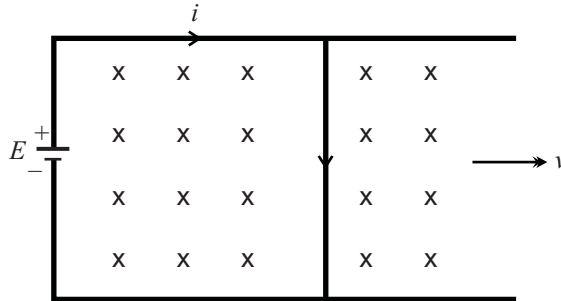


Figure 13-10: A set-up with a straight conducting wire lying between two fixed conducting rails and a battery of EMF E ; the wire can slide over the rails set at a distance l apart, and the entire set-up lies in a magnetic field, with uniform intensity B perpendicular to the plane of the rails (i.e., the plane of the figure; the magnetic field lines point into this plane); the magnetic field exerts a force on the wire because of the current flowing in it; if this force is sufficient to overcome the limiting friction between the wire and the rails, the former moves with a steady velocity; the power delivered by the source of EMF equals the rate at which work is done against the force of friction while keeping the wire moving, together with the rate of energy dissipation in the circuit, the latter given by Joule's law of heating; the principles underlying the operation of this set-up are analogous to those in a DC motor.

13.2.6 Self-inductance

I introduced the concept of magnetic flux linked with a closed circuit in a magnetic field or, more generally, with any closed contour in the field, in section 13.2.1. Consider a coil of wire carrying a current I . This current will produce a magnetic field of its own, and this field will result in a flux linked with the coil itself. What can one say about the flux linked with a closed circuit like a coil of wire due to the field produced by the current flowing in the circuit itself? Since the flux is a surface integral of the magnetic field intensity and the field intensity in turn is linearly related to the current, the most general statement that can be made is that the flux will be proportional to the current:

$$\Phi \propto I, \quad (13-16a)$$

or,

$$\Phi = LI, \quad (13-16b)$$

where L is a constant for the coil or the circuit under consideration. It depends on the shape and size of the closed path along which the current flows. For a circular coil of wire, for instance, it depends on the radius of the coil and the number of turns (N) in it, the dependence on N being one of proportionality if the turns of the coil are close to one another. This is because of the fact that the flux for a closely wound coil of N turns is N times the flux associated with a single turn.

The constant L is termed the *self-inductance* (at times referred to as, simply, the *inductance*) of the circuit through which the current flows, which can thus be defined as the flux linked with the circuit per unit current flowing through it. The unit of L is $\text{Wb} \cdot \text{A}^{-1}$, alternatively termed the *henry* (H).

The flux in equations (13-16a), (13-16b) is commonly defined to be the one with the unit vector \hat{n} in eq. (13-2) chosen in accordance with the right hand rule with reference to the sense of flow of the current in the coil (or current loop) under consideration. Thus, for a coil of several turns, the flux depends on the relative directions of flow of the current

through the successive turns, in which case the contribution to the flux due to some of the turns may get canceled by the contribution coming from the others.

13.2.6.1 Self-inductance of a long solenoid

We came across the expression for the magnetic intensity due to a long tightly wound solenoid in section 12.8.9.1, where it was seen that the field is, to a good degree of approximation, *uniform* throughout the interior of the solenoid. If the area of cross-section of the solenoid be A , then the flux linked with each turn of the solenoid winding is BA , where B is given by $B = \mu_0 n_0 I$ (check this out; refer to eq. (12-61c)). Here we assume the medium in the interior of the solenoid to be vacuum (air-cored solenoid; recall that the permeability of air is, approximately, μ_0) while, in practice, the coil of a solenoid is often wound on a magnetic material for which the permeability μ has a large value. The flux associated with the entire solenoid is then

$$\Phi = n_0 l B A, \quad (13-17a)$$

where l stands for the length of the solenoid (and thus $n_0 l$ is the total number of turns in it). Using the above expression for B and the defining equation, eq. (13-16b), the self inductance of the solenoid is seen to be

$$L = \mu_0 n_0^2 V, \quad (13-17b)$$

where $V = Al$ is the volume of the solenoid and n_0 is the number of turns per unit length, as in eq. (12-61c).

If the interior of the solenoid is filled with a material of permeability μ , then the expression for the self-inductance of the solenoid gets modified to

$$L = \mu n_0^2 V. \quad (13-17c)$$

Problem 13-6

Check out eq. (13-17c).

Answer to Problem 13-6

HINT: According to Ampere's circuital law (section 12.8.10), the expression for B gets modified to $B = \mu_0 I$.

13.2.6.2 Self-inductance of a toroidal solenoid

A toroidal solenoid is a tightly wound coil on a ring (fig. 13-11), commonly made up of a magnetic material (the winding being insulated from the material of the ring). The magnetic field intensity in the interior of toroid is once again given by the expression $\mu n_0 I$, where n_0 is the number of turns of the winding per unit length of the circumference of the ring, the radius of cross section of the ring being assumed to be small compared to the radius of its axial circumference. This is seen by imagining a closed path in the toroidal solenoid along its circular axis and employing the circuital law (check this out). Thus the self inductance is once again given by the expression (13-17c), where V is now the volume of the interior of the ring.

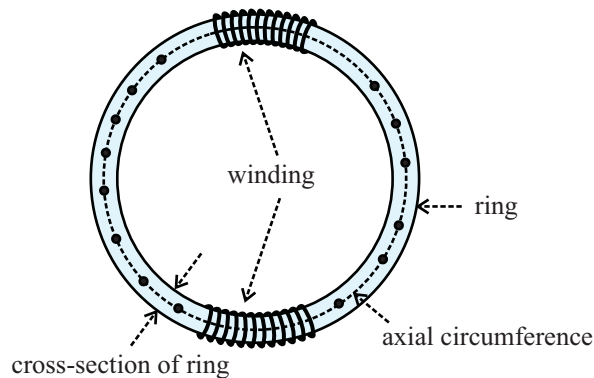


Figure 13-11: The toroidal solenoid; only parts of the winding on the solenoid are shown, the rest being represented by dots; the axial circumference of the toroid is assumed to be large compared to the circumference of the cross-section of the ring.

13.2.6.3 Inductor

A coil of wire (air-cored or wound on a core of magnetic material) that enhances the self inductance of a closed circuit when inserted in it, is referred to as an *inductor*. Fig. 13-12 (A) depicts a coil of wire (C) in a circuit made of a wire loop (W) and a DC source of EMF, the latter depicted with the help of its circuit symbol. Fig. 13-12(B) is a symbolic representation of the set-up, where the circuit symbol for the coil, functioning as an inductor, is shown along with that for the wire loop functioning as a resistor. In this figure L , E , and R stand respectively for the self-inductance of the inductor, the EMF of the DC source, and the resistance of the wire loop, including the resistance of the coil and the internal resistance of the source of EMF. To be precise, any conducting part of a circuit that can carry a current is endowed with a self inductance though, in practice, only those parts that have an appreciable value of the self inductance are referred to as inductors. At times, an inductor with inductance, say, L is referred to, simply, an inductance L for the sake of convenience (similarly, a resistor with resistance R is commonly referred to as a resistance R , and a capacitor with capacitance C is referred to as a capacitance C).

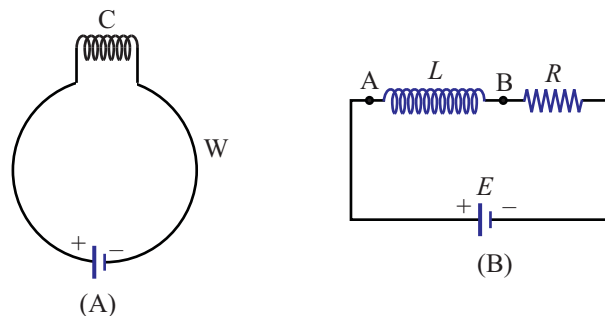


Figure 13-12: (A) depicting a coil (C) of wire inserted in a closed circuit made of a wire loop (W) and a DC source of EMF; (B) a representation of the set-up in terms of circuit symbols of the various elements; the symbol between the points A and B represents an inductor (inductance L), while the total resistance (R) in the circuit is also represented by the circuit symbol for a resistor; E stands for the EMF in the circuit.

13.2.6.4 Back EMF in an inductor

The self-inductance of an inductor (of inductance, say, L) relates to the magnetic flux linked with it due to a current (say, I) flowing through the inductor itself, the flux being given by the expression $\Phi = LI$. If, then, the current is made to vary with time, the flux also changes and an EMF is induced in the inductor. According to Faraday's law, the induced EMF is given by

$$E_{\text{induced}} = -\frac{d\Phi}{dt} = -L\frac{dI}{dt}. \quad (13-18)$$

The appearance of the negative sign in this expression has already been commented upon in section 13.2.2 (see, in particular, sec. 13.2.2.1). It tells us that the induced EMF acts in such a direction as to oppose the variation of the current. It is referred to as the *back EMF* in the circuit. If the circuit includes an external source of EMF, say, E , as in fig. 13-12 (A), (B), then the *net* EMF acting in the circuit will be

$$\mathcal{E} = E - L\frac{dI}{dt}. \quad (13-19)$$

13.2.7 Mutual inductance

Consider now a pair of coils carrying currents, say, I_1 , I_2 , where the coils may or may not belong to the same circuit. The magnetic field produced by the current (I_1) in the first coil results in a flux linkage (say, Φ_2) with the second coil, where Φ_2 is proportional to I_1 (a consequence of the principle of superposition pertaining to the magnetic field intensity in a magnetic field and the source currents producing the field). In a similar manner, the current I_2 in the second coil results in a flux linkage Φ_1 with the first coil where Φ_1 is proportional to I_2 . One can then write

$$\Phi_1 = M_{12}I_2, \quad \Phi_2 = M_{21}I_1, \quad (13-20)$$

where M_{12} and M_{21} are constants depending on the shapes and sizes of the coils, including the number of turns in each coil, and on the mutual disposition of the coils. A consideration of the energy required to set up the currents I_1 and I_2 in the two coils

leads to the result

$$M_{12} = M_{21} = M \text{ (say)}, \quad (13-21)$$

where the constant M is termed the *mutual inductance* of the two coils (or, more generally, of two circuits). The unit of mutual inductance in the SI system is, once again, the henry (H).

Formulae (13-20), (13-21) express the flux linkage with either of the two coils resulting from the current in the *other* coil. In reality, the current in each coil results in a flux linkage with the *same* coil as well, depending on its self-inductance. Thus, if L_1 and L_2 be the self-inductances of the two coils under consideration, the flux linked with these will be given by

$$\Phi_1 = L_1 I_1 + M I_2, \quad \Phi_2 = L_2 I_2 + M I_1. \quad (13-22)$$

While L_1 and L_2 are positive quantities, M can be either positive or negative depending on the relative directions of the currents in the two coils. For instance, with given directions of flow of the currents I_1 and I_2 in the two coils, the flux (Φ_1) in the first coil due to the current (I_2) in the second coil may cancel part of the flux in it due to its own current (I_1), in which case M will be a negative quantity. However, energy considerations show that the three quantities L_1 , L_2 , and M have to satisfy the inequality

$$L_1 L_2 \geq M^2. \quad (13-23)$$

The ratio $\frac{M^2}{L_1 L_2}$ tells us how effective the current in either of the coils is in producing a flux linkage with the other coil as compared to the flux produced by a current in that coil itself, and is termed the *coefficient of coupling* between the two coils.

If the currents I_1 and I_2 in the coils are made to vary with time, then there will be produced an induced EMF in each coil in accordance with Faraday's law, which will now be given by $-\frac{d\Phi_1}{dt}$ and $-\frac{d\Phi_2}{dt}$ as worked out from equations (13-22).

Problem 13-7

Consider a coil of N turns in the form of a square of side a , placed beside a long straight wire as shown in fig. 13-13, where the wire lies in the same plane as the coil. The arm of the coil closer to the wire is at a distance b from it. Find the mutual inductance M between the wire and the coil.

Answer to Problem 13-7

Imagine a current i to be flowing through the straight wire. This produces a magnetic field of strength $B = \frac{\mu_0 i}{2\pi x}$ at any point at a distance x from the wire. Choosing Cartesian axes parallel to the sides of the coil (with origin O on the straight wire and the x -axis passing through the center of the coil), and imagining an area $dA = dx dy$ (not shown in the figure) in it at a distance x from the wire, the flux through the area is $d\phi = NBdA = N \frac{\mu_0 i}{2\pi} \frac{dx}{x} dy$. Integrating over the area of the coil, the mutual inductance is found to be $M = \frac{\int d\phi}{i} = N \frac{\mu_0 a}{2\pi} \ln \frac{a+b}{b}$.

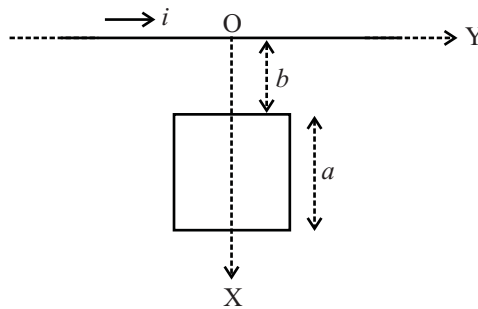


Figure 13-13: A long straight wire and a coil of N turns in the shape of a square, both lying in the same plane. The wire is parallel to two of the sides of the coil. The flux linked with the coil due to a current i in the straight wire is given by the expression $\phi = \frac{\mu_0}{2\pi} Ni \int_b^{a+b} \frac{dx}{x} \int_{-\frac{a}{2}}^{\frac{a}{2}} dy$, from which the mutual inductance $M = \frac{\phi}{i}$ between the wire and the coil can be worked out, as in problem 13-7.

13.3 Varying currents in electrical circuits

Imagine a conducting wire with its two ends connected between the terminals of an electrical cell as in fig. 12-6. As we saw in section 12.5.3 (eq. (12-19d)), a stationary state will prevail in the circuit, with a steady current $I = \frac{\mathcal{E}}{r+R}$ flowing through it. However, the magnetic field set up by the current through the wire can alter things a bit. As the

circuit is completed, the magnetic field intensity changes from a zero to a non-zero value, and this changing magnetic field induces a back EMF in the circuit. In accordance with Lenz's law, this back EMF tends to oppose the process causing the change in magnetic flux, i.e., in other words, tends to slow down the process of the current in the circuit changing from a zero value to the value I given above.

What this means is that the current cannot jump instantaneously from 0 to its final value I because of the self induction effect in the circuit itself, and it will take a certain time for the current to *grow* from 0 to I . One can thus speak of two stages or periods relating to the current in the circuit - a period of *varying* current when the current grows from zero value up to a value determined in accordance with Ohm's law, followed by the *steady* condition when the current remains constant at the value I . A similar phenomenon involving a varying current is observed as a circuit carrying a steady current I is *switched off*, when it is found that the current does not instantaneously jump from I to a zero value, but takes a certain time to *decay*. It is this intermediate stage of varying current in a circuit, during its growth or decay, that we will be interested in in the present section.

As we will see below, the time interval during which the current grows from 0 to I or decays from I to 0, may be short or long, depending on the self inductance in the circuit. Another instance in which one encounters a varying current in a circuit is one where a *capacitor* is included in it. Once again, the time interval during which the current varies depends on the value of the capacitance in the circuit. Finally, the way the current varies in a circuit including an inductor *and* a capacitor in addition to a resistor, will be seen to present interesting features where the variation can be either monotonic or *oscillatory*.

1. Strictly speaking, it takes an *infinite* time for the current to grow from 0 to I or to decay from I to 0. However, one can identify a certain finite characteristic time in which the process of growth or decay of current can, for practical purposes, be said to be completed. After the lapse of this characteristic time the current may be assumed to have become steady since one can then ignore its variation as being

negligible.

2. According to what has been stated above, a varying current in a circuit corresponds to an intermediate stage between two steady values (say, 0 and I). This is to be distinguished from an *alternating current*, to be discussed in subsequent sections below (AC generators and motors have already been introduced above). While the current in an AC circuit also changes with time, it is distinguished by the feature that it varies periodically for an indefinite length of time (as long, that is, as the circuit under consideration remains closed), i.e., its *variation is steady*, recurring in an identical pattern at certain fixed intervals of time. The latter are determined by the *time period* of the AC current which can be different in different situations.

13.3.1 Currents and voltages in an L - R circuit

13.3.1.1 Growth of current

Figure 13-14(A) depicts a circuit made up of a source of EMF, say, E connected across a combination of an inductor coil and a resistor. In an actual circuit the inductance L includes not only the self inductance of the coil but the self inductance due to other parts of the circuit as well though in practice the latter may be negligible in comparison with the inductance of the coil itself. Likewise, R stands for the total resistance of the circuit, including the internal resistance of the source of EMF and the resistance of the inductor coil. The circuit in fig. 13-14 contains a key which is shown in the closed configuration. Fig. 13-14(B) shows the open and closed configurations of the key.

Suppose that the circuit is closed at time $t = 0$, prior to which the current in it was zero. As the current builds up, let i denote the current at any later time t . As we saw in section 13.2.6.4 (refer to eq. (13-18)), the back EMF in the circuit at this point of time will be $-L \frac{di}{dt}$, where $\frac{di}{dt}$ stands for the rate of change of the current at time t . Hence the net EMF at time t is

$$\mathcal{E} = E - L \frac{di}{dt}. \quad (13-24)$$

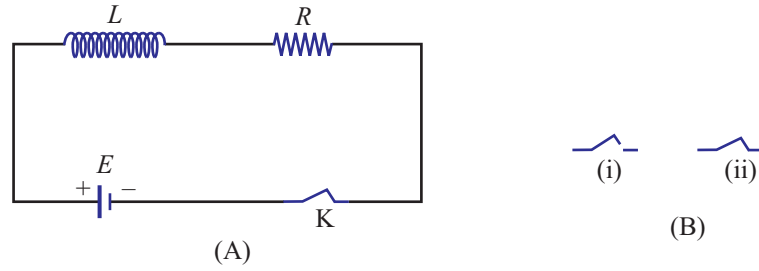


Figure 13-14: (A) An L - R circuit, with a DC source of EMF E and a key K ; the key is shown closed, allowing a current to flow in the circuit; (B) two configurations of the key - (i) open, and (ii) closed; as the key is closed in (A), the current in the circuit starts growing till, after a sufficiently long interval of time, the current attains a steady value.

Making use, then, of eq. (12-19d) and recognizing that, in fig. 13-14(A), R includes the internal resistance of the source of EMF, one arrives at

$$E - L \frac{di}{dt} = Ri, \quad (13-25a)$$

or,

$$\frac{di}{dt} + \frac{R}{L}i = \frac{E}{L}. \quad (13-25b)$$

This a *differential equation* relating the current i and its rate of change $\frac{di}{dt}$ at any given time t , which one can solve to find the current i as a function of t . For this, one requires the *initial condition* on the current which in the present instance is

$$i = 0 \text{ at } t = 0. \quad (13-25c)$$

The solution that comes out is

$$i(t) = \frac{E}{R} (1 - e^{-\frac{R}{L}t}). \quad (13-26)$$

It is straightforward to check that $i(t)$ as given in eq. (13-26) does indeed satisfy the differential equation (13-25b) along with the initial condition (13-25c).

One way to arrive at the solution (13-26) is to substitute $i' = i - \frac{E}{R}$ in eq. (13-25b),

wherein the latter can be written in the form

$$\frac{di'}{i'} = -\frac{R}{L}dt,$$

and then integrating from $i' = -\frac{E}{R}$ at $t = 0$ up to $i' = i(t) - \frac{E}{R}$ at time t . Incidentally, symbols like di' or dt in the above equation, written in isolation, do not have precisely defined meanings and are to be seen as nothing more than convenient shorthands. They are meaningful only under an integral sign.

Figure 13-15 depicts graphically the variation of current i with time t implied by eq. (13-26). Starting from $i = 0$, the current keeps on growing, ultimately reaching the value $I = \frac{E}{R}$ for $t \rightarrow \infty$. The rapidity with which the current grows is determined by the parameter $\tau = \frac{L}{R}$, termed the *time constant* of the L - R circuit under consideration. Fig. 13-15 shows the growth curves for two different values (τ_1 and τ_2 , with $\tau_2 > \tau_1$) of the time constant where it is seen that the current grows more slowly for τ_2 , the larger of the two time constants. In other words, the larger the value of self inductance (for any given resistance R), the slower is the growth of current. This is understandable since a larger self inductance means a larger back EMF, and hence a greater opposition to the growth of current.

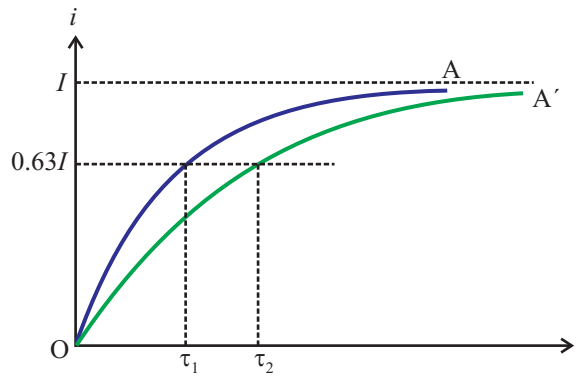


Figure 13-15: Growth of current (schematic) in L - R circuit with two different values (τ_1 , τ_2) of the time constant ($\tau_2 > \tau_1$); the current grows more slowly for the circuit with the larger value of the time constant; the points A and A' indicate that the current is very close to its final value I .

The current ultimately grows to the value $I = \frac{E}{R}$, which is independent of the self in-

ductance L in the circuit, i.e., the value dictated by Ohm's law. What also transpires from the growth curves in fig. 13-15 is that, though from a mathematical point of view, it takes an infinite time for the current to reach this final value, in reality the current grows to a value close to this in a *finite* time after the circuit is closed. For instance the current at either of the points A and A' in the two growth curves in fig. 13-15 can, to all intents and purposes, be assumed to be I . Indeed the current reaches a level of nearly 63% of its final value in a time τ , the time constant of the circuit. It is in this sense that τ can be taken as a measure of the time interval in which the current gets close to its final, steady value.

13.3.1.2 Decay of current

While I have described above the *growth* of current in an L - R circuit, one could equally well look at the *decay* of current in such a circuit from any given initial value to zero. Imagine that a steady current i_0 is established in a closed circuit made up of an inductance L and a resistance R with the help of a DC source of EMF appropriately connected to the inductance and the resistance (a parallel connection of the source may be necessary for this) and suppose that, at time $t = 0$, the DC source is removed so that now the current in the closed circuit made up of L and R decreases to zero. Here again, the current does not decrease to zero value all of a sudden, but there occurs a gradual decay, which may be more or less rapid depending on the values of L and R .

The differential equation governing the decay of current is seen to be

$$\frac{di}{dt} + \frac{R}{L}i = 0, \quad (13-27)$$

whose solution, subject to the boundary condition $i = i_0$ at $t = 0$ is

$$i = i_0 e^{-\frac{R}{L}t}. \quad (13-28)$$

Once again, it is the time constant $\tau = \frac{L}{R}$ that determines how rapidly the current decays to its final zero value.

13.3.2 Analysis of circuits with varying currents

The growth and decay of currents in L - R circuits analyzed above provide simple instances of *varying currents*. As we will see (refer to sec. 13.3.3), circuits involving capacitors constitute other instances of those with varying currents. In general, the analysis of circuits with varying currents follows principles analogous to those adopted in the analysis of DC circuits in sec. 12.7. In the case of DC circuits, the Kirchhoff principles were seen to be consequences of the principle of conservation of charge and of the principle of conservation of energy, where the latter implies that the sum of potential drops (including internal drops) across various arms in a closed mesh has to equal the net EMF in that mesh. Essentially the same principles can be made use of in the analysis of circuits with varying currents by taking into account the following additional considerations :

1. In a circuit in which no capacitance is included, there cannot be any accumulation of charge anywhere, and hence Kirchhoff first principle is applicable in respect of the instantaneous currents meeting at any junction. If, on the other hand, a terminal of a capacitor is connected to a junction, then the algebraic sum of the instantaneous currents meeting at that junction has to be equal to the rate of accumulation of charge at the capacitor terminal (a 'terminal' of a capacitor refers to one of the two conducting bodies constituting the capacitor).
2. The algebraic sum of potential drops across various arms of a closed mesh has to be equal to the net EMF in that mesh, where the net EMF is to be calculated by taking into account the back EMF's across the inductors that may possibly be included in the mesh.

Problem 13-8

Consider the circuit of fig. 13-16, with resistances R_1 , R_2 , r , inductance L , and a battery of EMF E , where r includes the resistance of the inductor. Find the expression for the current i at a time t after the key K is closed.

Answer to Problem 13-8

HINT: Referring to fig. 13-16, let the current through the branch containing L and r at any given time be i' , which means that the current through R_2 is $i - i'$ (Kirchhoff's first principle; no accumulation of charge takes place since there is no capacitor in the circuit). Kirchhoff's second principle, applied to the loop containing E , R_1 , and R_2 gives $i' = (1 + \frac{R_1}{R_2})i - \frac{E}{R_2}$. Next, consider the loop made up of L , r , and R_2 . A back EMF $L \frac{di'}{dt}$ acts in this loop (in opposition to the growth of i'), as a result of which one has $L \frac{di'}{dt} + i'r - (i - i')R_2 = 0$. Combining with the above expression for i' , one gets

$$\frac{dI}{dt} + \frac{r'}{L}I = 0,$$

where $I = i - \frac{E}{R_1 + \frac{rR_2}{r+R_2}} = i - i_0$ (say), and $r' = r + \frac{R_1R_2}{R_1+R_2}$. Evidently, the current i at $t = 0$, i.e., immediately after closing the circuit, is 0. Making use of this initial condition in solving the differential equation for I (and hence i), one gets $i = i_0(1 - e^{-\frac{r'}{L}t})$ (check this out). Thus, the time constant for the growth of current in the circuit is $\tau = \frac{L}{r'} = \frac{L}{r + \frac{R_1R_2}{R_1+R_2}}$, and the current at $t \rightarrow \infty$ (steady state) is $i_0 = \frac{E}{R_1 + \frac{rR_2}{r+R_2}}$, as it should be (why?).

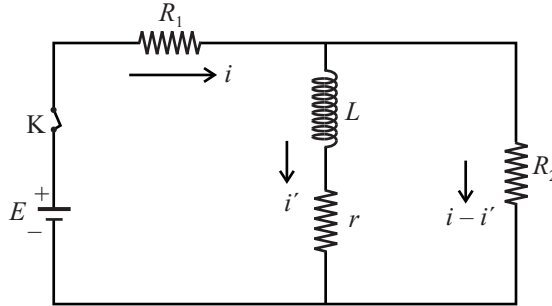


Figure 13-16: A circuit made up of resistances R_1 , R_2 , r , an inductance L , and a battery of EMF E , where r includes the resistance of the inductor; i and i' are the currents through R_1 and r respectively at a time t from the instant of closing the key K ; the current i varies with t as $i = i_0(1 - e^{-\frac{t}{\tau}})$, where $i_0 = \frac{E}{R_1 + \frac{rR_2}{r+R_2}}$, and the effective time constant is $\tau = \frac{L}{r + \frac{R_1R_2}{R_1+R_2}}$, as seen in problem 13-8.

Problem 13-9
Growth of currents in a coupled circuit

Fig. 13-17 shows a pair of L - R circuits coupled together, one with a DC source of EMF E_1 , an

inductance L_1 , and a resistance R_1 , and the other with corresponding values E_2 , L_2 , R_2 . The coupling occurs through the mutual inductance M between the circuits. Assuming that the currents $I_1(t)$, $I_2(t)$ in the two circuits are zero at $t = 0$, find the time constants describing the subsequent growth of currents in the two circuits.

Answer to Problem 13-9

The flux linked with the first circuit at any given time t is $L_1 I_1(t) + M I_2(t)$, which means that the back EMF at time t is $-(L_1 \frac{dI_1}{dt} + M \frac{dI_2}{dt})$. This gives the differential equation $E_1 - (L_1 \frac{dI_1}{dt} + M \frac{dI_2}{dt}) = R_1 I_1$ for the first circuit. Likewise, the differential equation for the second circuit is $E_2 - (L_2 \frac{dI_2}{dt} + M \frac{dI_1}{dt}) = R_2 I_2$. The solution to these differential equations satisfying the initial conditions at $t = 0$ is of the form

$$I_k(t) = A_k(1 - e^{-\lambda_1 t}) + B_k(1 - e^{-\lambda_2 t}) \quad (k = 1, 2),$$

where the constants λ_k , A_k , B_k ($k = 1, 2$) are to be obtained by substitution in the two differential equations. On substitution, each of the two differential equations yields three equations relating the unknown parameters: one resulting from the time independent terms, one from terms proportional to $e^{-\lambda_1 t}$, and one from terms proportional to $e^{-\lambda_2 t}$. Among the three pairs of equations so obtained, two pairs yield values of the constants λ_1 , λ_2 on consideration of the conditions for the existence of nontrivial solutions. The parameters A_k , B_k ($k = 1, 2$) are then obtained on making use of the values of λ_1 , λ_2 .

On following this procedure, one obtains

$$\lambda_{1,2} = \frac{1}{2(L_1 L_2 - M^2)} [(R_1 L_2 + R_2 L_1) \pm \sqrt{(R_1 L_2 - R_2 L_1)^2 + 4R_1 R_2 M^2}],$$

the corresponding time constants for the growth of currents in the circuits being λ_1^{-1} , λ_2^{-1} . The values of the remaining parameters are not relevant for our purpose.

In the limit $M \rightarrow 0$, the two circuits become uncoupled and the time constants reduce to $\tau_1 = \lambda_1^{-1} = \frac{L_1}{R_1}$, $\tau_2 = \lambda_2^{-1} = \frac{L_2}{R_2}$, as they should. In this case one obtains, in accordance with the initial conditions, $B_1 = 0$, $A_2 = 0$, which means that the currents grow in the two circuits independently of each other, with their respective time constants.

The limit $M^2 \rightarrow L_1 L_2$ is interesting in that, in this limit of maximum coupling between the two

circuits, λ_1 goes to infinity while λ_2 attains the value $\lambda_2 = \frac{R_1 R_2}{R_1 L_2 + R_2 L_1}$. In this case the currents first grow discontinuously in the two circuits and then undergo a continuous growth with a common time constant λ_2^{-1} .

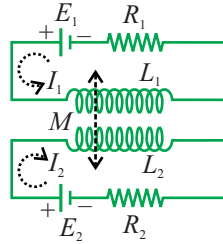


Figure 13-17: Growth of currents in a coupled circuit (problem 13-9); each of the circuits is made up of a DC source, an inductance, and a resistance, while the two are coupled by means of the mutual inductance M ; assuming that the currents $I_1(t)$, $I_2(t)$ are zero at $t = 0$ when the two circuits are closed, the current in either circuit grows subsequently with *two* time constants λ_1 , λ_2 because of the coupling; in the case $M \rightarrow 0$ the currents in the two circuits grow independently of each other, each with its own time constant; in the limit $M^2 \rightarrow L_1 L_2$, when the maximum possible coupling is achieved, one of the two time constants goes to infinity and the currents grow discontinuously at $t = 0$, whereafter the two currents grow with a common time constant.

13.3.3 Currents and voltages in a C - R circuit

13.3.3.1 Growth of charge

Fig. 13-18 depicts a circuit made up of a source of EMF (E) together with a capacitor of capacitance C and a resistor. The resistance R shown in the figure includes, in addition to the resistance of the resistor, the internal resistance of the source of EMF as well. Supposing that the capacitor was uncharged prior to completing the connection of the circuit, say, at time $t = 0$, there begins a flow of charge (in the clockwise sense in the figure) when the circuit is completed, wherein a charge accumulates on the capacitor.

Assuming the charge on the capacitor at time t to be q , the potential difference across the capacitor is seen to be $v = \frac{q}{C}$, which acts in opposition to the EMF E . The potential

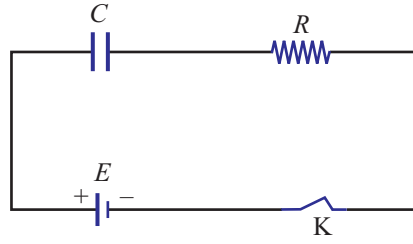


Figure 13-18: A C - R circuit connected to a DC source of EMF E ; on closing the key K , the charge in the capacitor grows, till the potential difference across the capacitor becomes E ; in the final, steady state, the current flowing through the circuit reduces to zero.

difference across the resistance at time t is then seen to be $E - \frac{q}{C}$, which gives

$$E - \frac{q}{C} = iR, \quad (13-29)$$

where i stands for the current in the circuit at time t . The current, however, is simply the rate at which charge accumulates on the capacitor, and is thus given by $i = \frac{dq}{dt}$ (reason this out). On differentiating both sides of eq. (13-29), then, we obtain the differential equation describing the way the current in the circuit changes with time:

$$\frac{di}{dt} + \frac{1}{CR}i = 0. \quad (13-30)$$

Referring to eq. (13-29), the initial condition in i in this case is seen to be $i = \frac{E}{R}$ at $t = 0$ since the charge q on the capacitor is initially zero. On solving the differential equation (13-30) along with this initial condition one finds

$$i(t) = \frac{E}{R}e^{-\frac{t}{CR}}. \quad (13-31)$$

Note that, for $t \rightarrow \infty$, one has $i \rightarrow 0$, i.e., the current *decays* to zero value over a large interval of time from the moment the circuit is closed. At the same time, the charge on the capacitor *grows* with time as (refer to eq. (13-29))

$$q(t) = EC(1 - e^{-\frac{t}{CR}}). \quad (13-32)$$

Fig. 13-19 depicts graphically the growth of charge in a C - R circuit from zero to the

final value $Q = EC$, which is independent of the resistance R in the circuit, which is expected since, for $t \rightarrow \infty$, the current in the circuit goes to zero, and hence the voltage drop across the resistance R is zero - the capacitor gets charged then to the *open circuit* voltage E . The rapidity with which the capacitor gets charged is determined by the product $\tau = CR$, the *time constant* of the C - R circuit. As the figure shows, when the growth curves for two different time constants (τ_1, τ_2 with, say, $\tau_2 > \tau_1$) are compared, the charge grows relatively more rapidly for the lower of the two time constants. Once again, the bulk of the charging (nearly 63%) is completed in a time interval equal to the time constant τ .

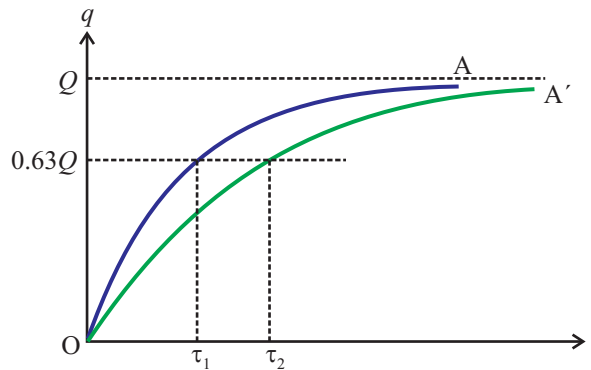


Figure 13-19: Growth of charge on the capacitor in a C - R circuit with two different values (τ_1, τ_2) of the time constant; the growth occurs relatively slowly for the larger value of the time constant; the charge effectively attains the final value Q after a large but finite time.

13.3.3.2 Decay of charge

Analogous to the decay of current in an L - R circuit, there occurs the decay of charge on a capacitor in a closed circuit made up of a capacitance C and a resistance R where, to start with, a charge q_0 has been given to the capacitor with the help of a DC source of EMF connected in parallel to both C and R . On disconnecting the source of EMF at time, say, $t = 0$, the capacitor will be seen to *discharge* through the resistor, with the charge on it ultimately going to zero.

The differential equation governing the decay of charge is seen to be

$$\frac{dq}{dt} + \frac{1}{RC}q = 0, \quad (13-33)$$

whose solution, subject to the boundary condition $q = q_0$ at $t = 0$ is

$$q = q_0 e^{-\frac{1}{RC}t}. \quad (13-34)$$

As in the case of the growth of charge, it is the time constant $\tau = RC$ that determines how rapidly the charge decays to its final zero value.

Problem 13-10

A capacitor of capacitance $C = 5.0 \mu\text{F}$ and a resistance R are connected in series with an ideal DC voltage source of EMF $E = 5.0 \text{ V}$. The potential difference across the resistor is found to be $V' = 3.0 \text{ V}$ after an interval of $t = 2.0 \text{ ms}$. Calculate the time constant of the combination and the value of R .

Answer to Problem 13-10

The potential difference across the capacitor at time t is given by $V(t) = E - V'$, and increases from 0 to E as $t \rightarrow \infty$. According to formula (13-32), $V(t)$ is given by $V(t) = E(1 - e^{-\frac{t}{CR}})$, i.e., $CR = \frac{t}{\ln(\frac{E}{E-V'})}$ (check this out). Making use of given values, one obtains the time constant $\tau = CR = 3.9 \text{ ms}$, and hence, $R = 0.78 \text{ k}\Omega$.

13.3.4 Oscillations in an L - C - R circuit

Finally, fig. 13-20 depicts a circuit containing an inductance (L), a resistance (R), and a capacitance (C) joined in series across a source of EMF (E). These may not, however, correspond to respective single elements in the circuit. For instance, R may include the resistance(s) of one or more resistors in the circuit along with the internal resistance of the source of EMF, as also the resistance(s) of one or more inductor coils.

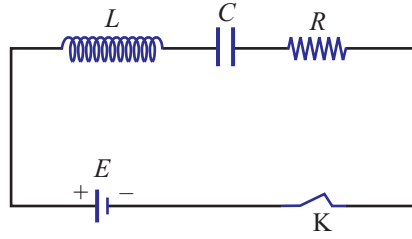


Figure 13-20: An L - C - R circuit with a DC source of EMF E , and a key K ; with the key closed, there occurs a damped oscillation of the current in the circuit; if, however, the resistance is sufficiently large, the variation of the current may be non-oscillatory.

The differential equation describing the variation of current i in the circuit is of the form

$$E - L \frac{di}{dt} - \frac{q}{C} = Ri, \quad (13-35a)$$

where q stands for the instantaneous charge on the capacitor, related to i as $i = \frac{dq}{dt}$ (check this out, following the approach in sections 13.3.1 and 13.3.3).

On differentiating with respect to t , the above equation gets transformed to

$$\frac{d^2i}{dt^2} + \frac{R}{L} \frac{di}{dt} + \frac{i}{CL} = 0. \quad (13-35b)$$

The same differential equation describes the variation of current in an L - C - R circuit when there is no source of EMF included in it, and a steady current (say, $i = i_0$) has been established in the circuit with the help of a DC source of EMF appropriately connected to it, where the DC source is *disconnected* at time, say, $t = 0$.

The equation (13-35b) is entirely analogous to the equation (4-41), with the parameters b and ω identified with $\frac{R}{2L}$ and $\frac{1}{\sqrt{LC}}$ respectively. Referring to the nature of solution of eq. (4-41), one can describe how the current i varies in the L - C - R circuit as well. Thus, for $\frac{R^2}{4L^2} < \frac{1}{LC}$, the variation of the current is oscillatory, with a gradually decaying amplitude, while for $\frac{R^2}{4L^2} > \frac{1}{LC}$, the current varies monotonically. These two conditions are said to correspond to *underdamped* and *overdamped* L - C - R circuits respectively. The special case $\frac{R^2}{4L^2} = \frac{1}{LC}$ corresponds to a *critically damped* circuit where the variation

of current is somewhat analogous to that in an overdamped circuit.

The frequency of oscillation of the current in an underdamped circuit is $\omega' = \sqrt{(\omega^2 - b^2)} = \sqrt{(\frac{1}{LC} - \frac{R^2}{4L^2})}$ (see eq. (4-42b)) which is close to $\sqrt{\frac{1}{LC}}$, the *natural frequency* of the circuit if the resistance R in the circuit is small.

The value of the current i at any given instant t , obtained by solving eq. (13-35b), depends on the initial conditions such as the values of i and $\frac{di}{dt}$ at time $t = 0$. For instance, at $t = 0$, one may have $i = i_0$, a value established with help of a DC source which is then disconnected, and $\frac{di}{dt} = 0$. The solution for the current at any arbitrary time $t > 0$ is then given by

$$i = i_0 e^{-bt} \left(\cos(\omega' t) + \frac{b}{\omega'} \sin(\omega' t) \right), \quad (13-36)$$

where b and ω' have been identified above, and where it is assumed that the circuit is underdamped (check the above statement out).

13.4 Magnetic field energy

Consider a closed loop of conducting wire of self inductance L and resistance R carrying a current i . Suppose that the current is zero to start with, and it gets built up to the level i as the two ends of the wire are connected to the terminals of a DC source (say, an electrical cell) of EMF E . As we have seen in sec. 13.3.1, the current varies with time so as to attain its final value i only after a certain time interval (which, in principle, is an infinitely long one, though in practice it is of the order of the time constant of the L - R circuit).

If, at any instant t after the current is switched on from zero value, the current in the circuit is i' , then the power delivered by the source of EMF is Ei' , so that the total energy delivered by the source of EMF is $\int_0^\infty Ei' dt$. According to eq. (13-25a), one can write this

as

$$\text{total energy delivered} = \int_0^\infty \left(L \frac{di'}{dt} i' + Ri'^2 \right) dt. \quad (13-37)$$

In this expression, the second term ($\int_0^\infty Ri'^2 dt$) gives the total energy dissipated in the wire due to its resistance. The remaining term in the expression for total energy delivered from the source of EMF can then be interpreted as the energy necessary to set up the current by overcoming the back EMF in the circuit. It can also be equivalently interpreted as the energy stored up in the magnetic field set up by the current. The expression for this magnetic energy can be written as

$$U_M = \int_0^i Li' di' = \frac{1}{2} Li^2. \quad (13-38)$$

This magnetic energy can be expressed in an alternative form in terms of the magnetic field intensity B at various points in the magnetic field. Imagining a small volume δv in the field, the energy associated with this small volume element is given by $\frac{1}{2} \mu B^2 \delta v$ (see below; refer to section 14.4.6.1), where μ stands for the permeability of the medium in which the field is set up. The total magnetic energy can then be expressed as the sum of the energies associated with all these volume elements distributed throughout the field, which reduces to the integral

$$U_M = \int \frac{1}{2\mu} B^2 dv, \quad (13-39)$$

where the integral is to be evaluated over the whole of space occupied by the field.

While the expressions (13-38) and (13-39) look quite different, they represent the same physical quantity in two different terms. As an instance of their identity, consider a closely wound long solenoid, for which the inductance L is given by eq. (13-17c), and hence, the expression (13-38) assumes the form

$$U_M = \frac{1}{2} \mu n_0^2 V i^2. \quad (13-40)$$

The expression (13-39), on the other hand, can be evaluated by noting that the field due

to a closely wound long solenoid is uniform within the volume of the solenoid, while it is zero at exterior points. The field at an interior point is given by the expression (12-61c), in which one has to replace μ_0 with μ if the interior of the solenoid is filled with a material medium, μ being the permeability of the medium. The integral in eq. (13-39) is then seen to reduce to precisely the expression in eq. (13-40) (check this out).

The energy associated with a magnetic field per unit volume around any given point, i.e., the energy *density* at that point, is thus

$$u_M = \frac{B^2}{2\mu}. \quad (13-41)$$

The expression (13-38) can be generalized to a number of conductors carrying given currents. For instance, considering a pair of conducting wires carrying currents i_1 and i_2 , the expression for the energy required to set up the currents is given by

$$U_M = \frac{1}{2}L_1i_1^2 + Mi_1i_2 + \frac{1}{2}L_2i_2^2, \quad (13-42)$$

where L_1 , L_2 , and M denote the self inductances of the two conductors, and their mutual inductance respectively (see problem 13-11, where a derivation is outlined). Once again, this energy can be interpreted as the energy associated with the magnetic field set up by the currents in the two conductors, where the energy density of the magnetic field at any given point is given by eq. (13-41), B being the magnitude of the magnetic field intensity produced by the two currents, at the point under consideration.

The term ‘conductor’ here refers to a closed circuit made of one or more conducting materials, carrying a given current.

Problem 13-11

Establish the formula (13-42).

Answer to Problem 13-11

HINT: Imagine the currents in the conductors to be increased from zero up to their final values (i_1, i_2) in small steps. At any given instant t , let the currents be i'_1, i'_2 and let these be increased by di'_1, di'_2 in time dt . The back EMF impeding the growth of current in the first conductor is then $L_1 \frac{di'_1}{dt} + M \frac{di'_2}{dt}$, while that for the second conductor is $L_2 \frac{di'_2}{dt} + M \frac{di'_1}{dt}$. Hence the energy expended in effecting the increments di'_1, di'_2 in time dt is $(L_1 \frac{di'_1}{dt} + M \frac{di'_2}{dt})i'_1 dt + (L_2 \frac{di'_2}{dt} + M \frac{di'_1}{dt})i'_2 dt$ (reason this out). Integrating from the initial to the final time when the currents are i_1, i_2 , one obtains the total energy expended for setting up the magnetic field as

$$U_M = \int_{i'_1=0, i'_2=0}^{i'_1=i_1, i'_2=i_2} \left[\frac{1}{2} L_1 d(i_1'^2) + M d(i_1' i_2') + \frac{1}{2} L_2 d(i_2'^2) \right],$$

which yields expression (13-42).

Problem 13-12

A coil C of small area $A = 3.0 \times 10^{-4} \text{ m}^2$ and number of turns $n = 200$ is placed inside a long solenoid S at a point on its axis, the plane of C being perpendicular to the axis of S. The solenoid is of length $l = 1.5 \text{ m}$, has $N = 1500$ turns, and is wound on an insulating cylinder of radius $R = 0.05 \text{ m}$. When there is no current in the solenoid, a current $i = 1.5 \text{ A}$ in C, produces a flux $\Phi = 3.0 \times 10^{-6} \text{ Wb}$ linked with it. Calculate the self inductances of C and S, and the mutual inductance between the two. Supposing that, in addition to the current i in C, a current $I = 3.0 \text{ A}$ is set up in S. Obtain the magnetic energy stored in the system, assuming the mutual energy to be positive.

Answer to Problem 13-12

In accordance with formula (13-22), a current i in C, without any current flowing in S, results in a flux $\Phi = L_1 i$ linked with it (check this out), where L_1 is the self inductance of C. From the given data, one obtains $L_1 = \frac{\Phi}{i} = 2.0 \times 10^{-6} \text{ H}$. The current I in the long solenoid produces a magnetic field intensity of magnitude $B = \frac{N}{l} \mu_0 I$ in a direction parallel to the axis, and hence a flux $\phi = nAB$ linked with the coil C. This means that the mutual inductance of C and S is $M = \mu_0 \frac{N}{l} nA$. Making use of given values, one gets $M = 7.54 \times 10^{-5} \text{ H}$. An estimate for the flux linkage with S due to the current I in it is $\phi' = \pi R^2 N B$, i.e., its self inductance is $L_2 = \mu_0 \pi R^2 \frac{N^2}{l} = 1.48 \times 10^{-2} \text{ H}$. The magnetic energy of the system, with currents i and I in C and S respectively, is given by $U_M = \frac{1}{2} L_1 i^2 + \frac{1}{2} L_2 I^2 + M i I$. Using appropriate values of the various quantities in this expression,

one gets $U_M = 2.25 \times 10^{-6} + 6.66 \times 10^{-2} + 3.39 \times 10^{-4}$ J. One observes here that the energy is dominated by the self energy of the solenoid.

13.5 Alternating currents

13.5.1 Mathematical description of AC currents and voltages

Imagine that the current (I) in a circuit varies with time as

$$I = I_0 \cos(\omega t + \delta), \quad (13-43)$$

where I_0 , ω , and δ are constants whose significance I will presently indicate. Such a variation is referred to as a *sinusoidal* one, alternative forms of which are

$$I = I_0 \sin(\omega t + \delta), \quad I = I_1 \cos \omega t + I_2 \sin \omega t, \quad (13-44)$$

where I_0 , δ , I_1 and I_2 stand for new constants.

The first expression in (13-44) is obtained from the expression in (13-43) by the substitution $\delta \rightarrow -\frac{\pi}{2} + \delta$, and the second expression by $I_0 \cos \delta \rightarrow I_1$, $-I_0 \sin \delta \rightarrow I_2$.

Referring back to section 4.3, such a sinusoidal variation is an instance of a simple harmonic variation of a physical quantity, namely, the current in a circuit in the present context. A graphical plot of I against t will look similar to the ones in figures 4-3(A), (B).

One then says that the circuit under consideration carries an *alternating current* (AC in brief; though slightly anomalous, the terms *AC current* and AC voltage are in common use). Such a current may be set up by connecting, say, a resistor across the terminals of an AC generator (refer to sec. 13.2.3) or across an AC source of EMF generated by an electronic *oscillator* circuit. Similar to an AC current, other AC electrical quantities like an AC electromotive force or an AC voltage involve a sinusoidal variation with time.

Thus, an AC electromotive force varies with time as

$$E(t) = E_0 \cos(\omega t + \delta), \quad (13-45a)$$

and an AC voltage as

$$V = V_0 \cos(\omega t + \delta), \quad (13-45b)$$

where E_0 and V_0 are constants, as are ω and δ . Sources of AC and DC EMF are referred to as, simply, AC and DC sources respectively.

13.5.1.1 Amplitude, frequency, and phase

As explained in sections 4.1.1 and 4.3, the constant I_0 in the expression 13-43 and the first expression in (13-44) represents the *amplitude* of the alternating current (I_0 is commonly chosen to be positive), meaning that $I(t)$ varies with time in the range from $-I_0$ to I_0 . Of the other two constants in (13-43), ω stands for the angular frequency, related to the frequency (ν) and the time period (T) of variation of the current as $\omega = 2\pi\nu$ and $\omega = \frac{2\pi}{T}$ respectively. Finally, δ stands for the constant part of the *phase* $\Phi = \omega t + \delta$ or, in other words, the *initial* phase. Recall, in this context, the relation between the phase Φ and the phase *angle*, or the *reduced* phase ϕ , the latter being defined as $\phi = \Phi \bmod 2\pi$ (however, ϕ is also commonly referred to as the phase).

13.5.1.2 Root mean squared values

Consider an AC current in a circuit, given by the expression (13-43). The *squared* value of the current at time t is then

$$I(t)^2 = I_0^2 \cos^2(\omega t + \delta). \quad (13-46)$$

The *mean* value of this expression, calculated over a complete time period of the current (i.e., the *mean squared* current), is then

$$\langle I^2 \rangle = \frac{1}{T} \int_0^T I_0^2 \cos^2(\omega t + \delta) dt, \quad (13-47)$$

where the angular brackets are used to denote the mean value of a quantity.

The mean of a temporally varying quantity $f(t)$ is defined as

$$\langle f(t) \rangle = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \int_0^\tau f(t) dt, \quad (13-48)$$

provided that the limit exists. For a periodically varying quantity with time period T , this reduces to

$$\langle f(t) \rangle = \frac{1}{T} \int_0^T f(t) dt. \quad (13-49)$$

On evaluating the integral in eq. (13-47), one obtains

$$\langle I^2 \rangle = \frac{I_0^2}{2}, \quad (13-50)$$

(check this out).

The *square root* of the mean squared current is referred to as the *root mean squared* (RMS in brief) current. Denoting this by I_{RMS} , one has

$$I_{\text{RMS}} = \frac{I_0}{\sqrt{2}}. \quad (13-51a)$$

This is an important result: *the RMS value of an AC quantity is $\frac{1}{\sqrt{2}}$ times the amplitude of that quantity.*

Similar considerations relating to the AC voltage given by the expression (13-45b), lead to another instance of the above rule:

$$V_{\text{RMS}} = \frac{V_0}{\sqrt{2}}. \quad (13-51b)$$

Root mean squared values of currents and voltages are found to be relevant in considerations relating to *power* in an AC circuit (see sec. 13.5.5 below).

13.5.1.3 The complex representation of AC quantities

A very convenient way of representing sinusoidally varying quantities is by means of *complex numbers*. Recall that the expressions $a \cos \Phi$ and $a \sin \Phi$ are respectively the real and imaginary parts of the complex expression $ae^{i\Phi}$. Thus, one can write, for instance,

$$a \cos(\omega t + \delta) = \operatorname{Re}(Ae^{i\omega t}), \quad (13-52a)$$

where

$$A = ae^{i\delta}. \quad (13-52b)$$

In this mode of representation, A is referred to as the complex amplitude, to be distinguished from a , the real amplitude (or, simply, the *amplitude*; the complex amplitude is also referred to at times as just the amplitude); and δ is termed the phase angle of this complex amplitude. The expression $Ae^{i\omega t}$ is referred to as the *complex representation* of the sinusoidally varying quantity $a \cos(\omega t + \delta)$.

The complex representation of an AC quantity is commonly denoted with the same symbol as the quantity in question, with a tilde overhead. Thus, for instance, the complex representation of the current $I = I_0 \cos(\omega t + \delta)$ is written as $\tilde{I} = \tilde{I}_0 e^{i\omega t}$, where $\tilde{I}_0 = I_0 e^{i\delta}$, and one will have

$$I = \operatorname{Re} \tilde{I}, \quad I_0 = |\tilde{I}_0|. \quad (13-53)$$

In a similar manner, $\tilde{V} = \tilde{V}_0 e^{i\omega t}$ is the complex representation of $V = V_0 \cos(\omega t + \delta)$, where $\tilde{V}_0 = V_0 e^{i\delta}$, δ being once again the phase angle of \tilde{V}_0 .

In considering the AC currents and AC voltages in a given circuit, one can replace the real expressions for the currents and voltages by the corresponding complex expressions, and perform mathematical operations with these complex expressions. At any stage of these operations, one can revert to the real quantities by simply taking the real parts out of the complex ones (in the case of complex amplitudes, though, one

has to evaluate their magnitudes so as to get the corresponding real amplitudes). This use of complex representations is made possible by the fact that the relations between currents and voltages are all *linear* ones (nonlinear current-voltage relations are relevant in electronic circuits (see chapter 19) and will not be considered in this chapter). Great simplifications are thereby achieved in calculations leading to meaningful results relating to the circuit under consideration.

One such simplification results from the fact that, if the currents and voltages vary sinusoidally with time with an angular frequency ω , all relations between these quantities involve a common factor $e^{i\omega t}$ (an equivalent representation is also possible with $e^{-i\omega t}$ as the common factor; such a representation is commonly used in describing electromagnetic waves, see chapter 14), and one may then write these relations by canceling this common factor. The temporal variation then does not appear explicitly in the relations between the currents and voltages, which involve only the complex amplitudes of these quantities. This then leads one to a corresponding relation between the real amplitudes and the phase angles.

Examples of the use of complex representations of AC currents and voltages will be encountered in the following sections.

The complex representation of sinusoidally varying quantities finds its use not only in the description of AC currents and voltages, but in the description of time harmonic (i.e., *monochromatic*) fields as well. Examples of such fields arise in the context of waves such as acoustic waves and electromagnetic waves. The use of complex quantities in the description of various features of electromagnetic waves will be encountered in chapter 14 and, in the particular context of optics, in chapter 15.

While the complex representation makes use of the tilde on top of the relevant symbols denoting the various sinusoidally varying quantities, the tilde notation at times becomes cumbersome to look at. One then omits the tilde, with a mention of the intended meanings of the symbols whenever there is a possibility of confusion.

Sinusoidally varying quantities in their above complex forms are often referred to as *phasors*.

Problem 13-13

A circular coil of radius $a = 0.05\text{m}$ and of $N = 100$ turns, is placed in a region R in which a spatially uniform magnetic field is set up that varies sinusoidally with time, the field intensity being given by $\mathbf{B}(t) = \frac{1}{\sqrt{2}}B_0 \cos(\omega t + \delta)(\hat{j} + \hat{k})$ with amplitude $B_0 = 1.5 \times 10^{-2}\text{T}$, angular frequency $\omega = 600\text{s}^{-1}$, and initial phase $\delta = \frac{\pi}{6}$. Here \hat{j} , \hat{k} are unit vectors along the y- and z-axes of a right handed Cartesian co-ordinate system, the plane of the coil being parallel to the x-y plane of the co-ordinate system. Obtain the complex amplitude of the induced EMF in the coil.

Answer to Problem 13-13

HINT: The flux linked with the coil at time t is $\phi(t) = N\pi a^2 \hat{k} \cdot \mathbf{B}(t) = \frac{1}{\sqrt{2}}N\pi a^2 B_0 \cos(\omega t + \delta)$, where the perpendicular to the plane of the coil in the positive direction of the z-axis is chosen for the sake of reference. By Faraday's law of induction, the induced EMF at time t is $E(t) = -\frac{d\phi}{dt} = \frac{1}{\sqrt{2}}N\pi a^2 \omega B_0 \sin(\omega t + \delta)$, where the EMF acts in the sense related to the positive direction of the z-axis by the right hand rule. Denoting the complex expression for the induced EMF by $\tilde{E}(t)$, and making use of the defining relation $E(t) = \text{Re}\tilde{E}(t)$, one obtains $\tilde{E}(t) = [\frac{1}{\sqrt{2}i}N\pi a^2 \omega B_0 e^{i\delta}]e^{i\omega t}$. The complex amplitude of the induced EMF is then given by $\tilde{A} = \frac{1}{\sqrt{2}}N\pi a^2 \omega B_0 e^{i(\delta - \frac{\pi}{2})}$. The phase δ_0 is commonly chosen to lie in the interval $-\pi$ to π (an alternative choice is the interval from 0 to 2π). Hence, with $\delta = \frac{\pi}{6}$, the phase angle is to be taken as $\delta_0 = \frac{\pi}{6} - \frac{\pi}{2} = \frac{-\pi}{3}$. Substituting the given values of the other relevant parameters, one obtains $\tilde{A} = 5.0e^{\frac{-i\pi}{3}}\text{V}$.

13.5.2 An L - C - R circuit with an AC source

Fig. 13-21 shows an inductance (L), a capacitance (C), and a resistance (R) connected in series across an AC source of EMF $E(t) = E_0 \cos \omega t$. Once again, the inductance, capacitance, and resistance need not be lumped into single circuit elements but each of these may be distributed over several elements. The resistance R , for instance, includes the internal resistance of the AC source and the resistance of one or more inductor coils.

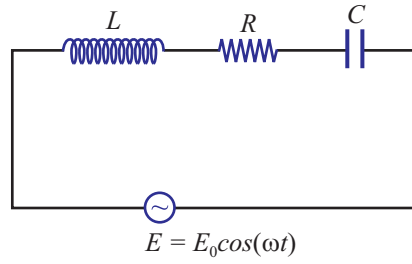


Figure 13-21: An L - C - R circuit with an AC source of EMF of angular frequency ω ; in the steady state, the current oscillates in the circuit with the angular frequency ω of the source of EMF.

This circuit is similar to the one of fig. 13-20, with the difference that it involves an AC source with a time-dependent EMF $E(t)$ instead of a DC source with a constant EMF E . Hence, the instantaneous current in the circuit and the charge in the capacitance at time t will be given by eq. (13-35a), with E replaced by $E_0 \cos \omega t$. On differentiating both sides of the resulting equation, equating $\frac{dq}{dt}$ to I (a small change in notation: the instantaneous current will now be denoted by I ; the symbol i will stand for the square root of -1), and rearranging terms, one arrives at

$$L \frac{d^2 I}{dt^2} + R \frac{dI}{dt} + \frac{I}{C} = \frac{d}{dt}(E_0 \cos \omega t). \quad (13-54)$$

The solution to this differential equation can, in general, be expressed as a sum of two parts, a *transient part* and a *steady part*, which is exactly similar to what we found in the periodically forced oscillations of a damped simple harmonic oscillator (see section 4.6). Indeed, as mentioned there, the equation of motion of a forced and damped simple harmonic oscillator (eq. (4-50)) has an exact correspondence with the equation (13-54). This correspondence becomes apparent on considering eq. (4-50) and eq. (13-54) in the complex representation and making the following substitutions:

$$x \rightarrow I, \quad m \rightarrow L, \quad \gamma \rightarrow R, \quad k \rightarrow \frac{1}{C}, \quad p \rightarrow \omega, \quad A \rightarrow iE_0\omega.$$

Thus, what we found for the transient part of the solution holds in the present context

as well. In particular, the transient part is nothing but the solution to eq. (13-35b) (which corresponds to eq. (4-39), describing a damped simple harmonic motion in the absence of a periodic forcing), and it becomes negligibly small at large times because of the damping effect of the resistance R in the L - C - R circuit. Thus, starting from the time the circuit is closed, if one waits for a sufficiently long time, one will find that the solution for $I(t)$ is described by the steady part alone.

This steady solution is obtained from the expression (4-52a) by appropriately making the above substitutions (the factor i in the last of the substitution formulas indicates that the phase is to be appropriately modified), which then represents a sinusoidally varying current with angular frequency ω . However, I am going to derive this steady solution by making use of the complex representation of AC currents and voltages outlined in sec. 13.5.1.3. Thus, we make the following replacements:

$$I(t) \rightarrow \tilde{I}_0 e^{i\omega t}; E(t) = E_0 \cos \omega t \rightarrow E_0 e^{i\omega t},$$

where \tilde{I}_0 stands for the complex amplitude of the current, while the complex amplitude of the EMF is, in the present instance, the same as its real amplitude E_0 , since the constant part of the phase in $E(t)$ has been assumed to be zero.

In a given circuit, one can assume the phase of a single current or voltage to be zero, since this leaves unaltered the *relative phases* of the various relevant quantities. The relative phases are the ones that determine physically observable features such as the power dissipation in any part of the circuit.

At the same time, we note that a differentiation with respect to t brings out a factor $i\omega$, since $\frac{d}{dt}e^{i\omega t} = i\omega e^{i\omega t}$. This leads us then to the following equation for \tilde{I}_0 , the complex amplitude of the steady state current in the circuit:

$$\left(-\omega^2 L + i\omega R + \frac{1}{C}\right)\tilde{I}_0 = i\omega E_0, \quad (13-55a)$$

from which one obtains

$$\tilde{I}_0 = \frac{E_0}{R + (i\omega L + \frac{1}{i\omega C})}. \quad (13-55b)$$

Let us write this as

$$\tilde{I}_0 = \frac{E_0}{\tilde{Z}}, \quad (13-56a)$$

where \tilde{Z} is referred to as the *complex impedance* of the circuit under consideration. It can be written in the form

$$\tilde{Z} = Ze^{i\theta}, \quad (13-56b)$$

where Z , the magnitude of the complex impedance (commonly termed, simply, the *impedance*; at times, the complex impedance itself is referred to, for the sake of brevity, as the impedance), and θ , the *phase angle* of the impedance, are given by

$$Z = (R^2 + (\omega L - \frac{1}{\omega C})^2)^{\frac{1}{2}}, \quad (13-56c)$$

$$\theta = \arctan \frac{\omega L - \frac{1}{\omega C}}{R}. \quad (13-56d)$$

The inverse tangent in eq. (13-56d) is commonly defined to lie in the range $-\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}$. With this definition, a negative value of θ means that the current in the AC circuit *leads* the EMF in phase while a positive value implies that the current *lags* behind in phase.

One can now recover the current $I(t)$ in the circuit at time t by taking the real part of $\tilde{I}e^{i\omega t}$, which gives

$$I(t) = \frac{E_0}{(R^2 + (\omega L - \frac{1}{\omega C})^2)^{\frac{1}{2}}} \cos(\omega t - \theta), \quad (13-57)$$

with θ given by eq. (13-56d).

I leave it to you to check that this expression for $I(t)$ is obtained from the steady state solution $x_0(t)$ for the forced oscillator, given by eq. (4-52a), with the substitution mentioned above (the factor i in the substitution $A \rightarrow i\omega E_0$ will require for some caution here).

This gives I_0 , the amplitude of the AC current in the circuit, as also its phase relative to the applied AC EMF. The expression for I_0 is seen to be

$$I_0 = \frac{E_0}{(R^2 + \frac{L}{C}(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega})^2)^{\frac{1}{2}}}, \quad (13-58)$$

where ω_0 is given by the expression (13-60) below. An alternative expression for I_0 in terms of the dimensionless parameters $\frac{\omega}{\omega_0}$ and $R\sqrt{\frac{C}{L}}$ is

$$I_0 = \frac{E_0}{R(1 + \frac{L}{CR^2}(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega})^2)^{\frac{1}{2}}}. \quad (13-59)$$

The variation of the amplitude with the frequency ω of the applied EMF for given values of L , C , and R is analogous to what we saw in section 4.6, and is shown in fig. 13-22. Similar to the response of a damped simple harmonic oscillator to an impressed periodic force, the response of the L - C - R circuit to an impressed alternating EMF involves the possibility of *resonance*, which corresponds to a large value of the amplitude I_0 of the current when the angular frequency of the alternating EMF is ω_0 (see eq. (13-58)) given by

$$\omega_0 = \frac{1}{\sqrt{LC}}, \quad (13-60)$$

the *resonant frequency* of the circuit.

I again leave it to you to check the above statement out and to show further that the maximum value of the amplitude I_0 as a function of ω becomes arbitrarily large as R is made to decrease towards zero value.

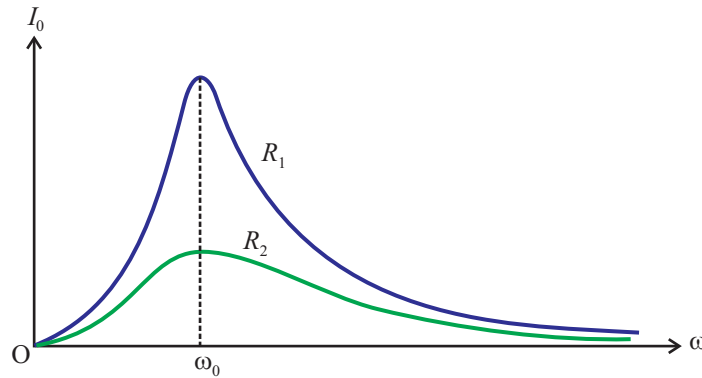


Figure 13-22: Resonance curve for the L - C - R circuit for two different values of $R\sqrt{\frac{C}{L}}$; the labels on the curves indicate the value of R ($R_2 > R_1$) for fixed L and C .

Note that, for sufficiently small values of R (more precisely, of $R\sqrt{\frac{C}{L}}$), the resonant frequency given by eq. (13-60) is the same as the frequency of oscillation of the current in an L - C - R circuit found in sec. 13.3.4, where the latter can be looked upon as the natural frequency of the circuit, analogous to the natural frequency (ω) of the oscillator of sections 4.5 and 4.6 (recall that the natural frequency gets changed only to a small extent when the damping constant is small).

Note finally that the phase difference θ between the impressed alternating EMF and the current in the L - C - R circuit is *zero* at resonance, when one has

$$I(t) = \frac{E_0}{R} \cos \omega t. \quad (13-61)$$

Thus, at resonance, the circuit behaves as if just the resistance R is connected across the AC source. In a sense, the effects of the inductance and the capacitance on the response of the circuit annul each other at resonance.

13.5.3 Impedance

The equation (13-56a), which relates the complex amplitude of the current to that of the impressed alternating EMF, is important in its own right, and completely describes the response of the circuit to the impressed EMF. It looks similar to the relation $I = \frac{E}{R}$ for a circuit in which a DC EMF is impressed on a resistance R . Thus, the complex impedance

\tilde{Z} of the AC circuit plays a role analogous to that of the resistance in a DC circuit in that it determines the response of the circuit to the EMF applied to it. However, as opposed to R , \tilde{Z} is a *complex* quantity, having a magnitude Z , which determines the real amplitude of the current, and a phase angle θ which determines the phase difference between the EMF and the current in the AC circuit.

For the L - C - R circuit considered in this section, the complex impedance \tilde{Z} is given by (refer to equations (13-55b) and (13-56a))

$$\tilde{Z} = R + i\omega L + \frac{1}{i\omega C}, \quad (13-62)$$

corresponding to which the (real) impedance Z and the phase angle θ are given by equations (13-56c) and (13-56d) respectively.

In the following, I introduce another (small) change in notation and denote the square root of -1 by j instead of i , in conformity with common usage in AC circuit theory (this will make possible the use of the symbol i to denote a current).

Making use of this new notation, one notes that, in addition to the form $\tilde{Z} = Ze^{j\theta}$, the complex impedance can be written in the form $\tilde{Z} = R + jX$ as well where, in the present context, X is given by $X = \omega L - \frac{1}{\omega C}$. Here, R and jX are referred to as the *resistive* and the *reactive* parts of the complex impedance respectively. The reactive part of a complex impedance arises, in general, due to the presence of inductors and capacitors in a circuit. The magnitude of the reactive part of impedance ($|X|$) is referred to as the *reactance* of the circuit at the angular frequency under consideration. The two representations, $\tilde{Z} = Ze^{j\theta}$, and $\tilde{Z} = R + jX$, are related to each other as

$$R = Z \cos \theta, X = Z \sin \theta. \quad (13-63)$$

A number of important corollaries follow from the above results. Suppose, for instance, that we have a circuit made of just an inductance L connected across an alternating source of EMF whose complex representation is $\tilde{E} = E_0 e^{j\omega t}$. With reference to the circuit

of fig. 13-21, this means that $R = 0$ and, in addition, $C \rightarrow \infty$ since now the last term on the left hand side in eq. (13-54) has to be zero.

Removing the resistance and the capacitance from the circuit of fig. 13-21 corresponds to a situation where the two ends of the resistance are at the same potential and so are the two ends of the capacitance. Thus the potential across the capacitance has to be zero, which is equivalent to taking $C \rightarrow \infty$.

This then gives, for a circuit made up of a single inductance L connected across an AC source,

$$\tilde{Z} = j\omega L, \quad (13-64a)$$

corresponding which the magnitude of the impedance (Z) and the phase angle are given by

$$Z = \omega L, \quad \theta = \frac{\pi}{2}. \quad (13-64b)$$

Thus, the impedance of an inductance is a purely reactive one, and one says that the *reactance* of an inductor with inductance L at an angular frequency ω is ωL . Moreover, the current in the inductance lags behind the EMF by a phase angle $\frac{\pi}{2}$.

Similar results are arrived at for a *capacitance* (C) connected across a source of alternating EMF. In this case, putting $L = 0$, $R = 0$, in eq. (13-62), one obtains

$$\tilde{Z} = -j\frac{1}{\omega C}, \quad Z = |X| = \frac{1}{\omega C}, \quad \theta = -\frac{\pi}{2}. \quad (13-65)$$

In other words, the impedance of a capacitor is a purely reactive one where the reactance at angular frequency ω is $\frac{1}{\omega C}$. Thus, the reactance of a capacitor becomes infinitely large for $\omega \rightarrow 0$, which corresponds to the limit of a DC EMF: the capacitor *blocks* the current when connected to a DC source. On the other hand, its reactance becomes negligibly small for very high frequencies. Finally, the current in the capacitor *leads* the

alternating EMF by the phase angle $\frac{\pi}{2}$.

The use of complex quantities relating to AC circuits allows us to write down rules analogous to those for DC circuits. One such rule relates the complex current \tilde{I} through an impedance with the complex potential difference \tilde{V} across it as

$$\tilde{V} = \tilde{I}\tilde{Z}. \quad (13-66)$$

Other rules of this kind relate to *series* and *parallel* combinations of impedances, to be discussed in section 13.5.4.

In section 13.5.2 I have presented an analysis of a *series L-C-R* circuit where an inductance, a capacitance, and a resistance are connected in series across an AC source. Analogous to DC circuits involving *networks* of resistors, AC circuits may also contain various possible combinations of series and parallel connections involving resistances, inductances, and capacitances. One can define an *equivalent impedance* for such a combination in a manner similar to the definition of equivalent resistance in a DC circuit. If the phase angle of this equivalent impedance is negative, the combination is said to have a *capacitive* reactance, while a positive phase angle corresponds to an *inductive* reactance.

13.5.4 Analysis of AC circuits

The series *L-C-R* circuit shown in fig. 13-21 can be looked upon as one containing a series combination of three impedances where the inductance, the capacitance, and the resistance are each considered as an impedance in its own right. Among these, the inductance and the capacitances are purely *reactive* impedances, with complex impedance $\tilde{Z}_L = j\omega L$ and $\tilde{Z}_C = -\frac{j}{\omega C}$, while the resistance is a purely *resistive* impedance characterized by $\tilde{Z}_R = R$.

Formula (13-62) then tells us that the impedance of the series combination of the three

components is simply the sum of the impedances considered severally:

$$\tilde{Z} = \tilde{Z}_R + \tilde{Z}_L + \tilde{Z}_C. \quad (13-67)$$

In other words, a series combination of a number of impedances is characterized by an *equivalent impedance* which is simply the sum of the impedances connected in series. This rule of series combination of impedances is analogous to the one relating to the equivalent resistance of a series combination of a number of resistances (see section 12.6.1). Indeed, the same logic as used in establishing the series rule for resistances applies in the present context as well, since the same fundamental principle operates in the two cases: the currents flowing through all the components connected in series are the same.

In a similar manner, the rule of *parallel* combination of a number of impedances in an AC circuit is also found to be analogous to the one we found in the case of parallel combination of a number of resistances, the basic operating principle here being, *the potential difference across each of the impedances connected in parallel are the same*.

Summarizing, the basic rules relating to the equivalent impedance of a series or parallel combination of impedances in an AC circuit are as follows:

$$\text{(series combination)} \quad \tilde{Z}_{\text{eq}} = \sum_{i=1}^N \tilde{Z}_i, \quad (13-68a)$$

$$\text{(parallel combination)} \quad \tilde{Z}_{\text{eq}}^{-1} = \sum_{i=1}^N \tilde{Z}_i^{-1}, \quad (13-68b)$$

where \tilde{Z}_{eq} stands for the equivalent complex impedance of N number of complex impedances \tilde{Z}_i ($i = 1, 2, \dots, N$) connected in series or in parallel respectively. Fig. 13-23(A), (B) illustrates the series and parallel combination respectively, of three impedances \tilde{Z}_1 , \tilde{Z}_2 , and \tilde{Z}_3 , corresponding to which the real impedances are, respectively, Z_1 , Z_2 , Z_3 .

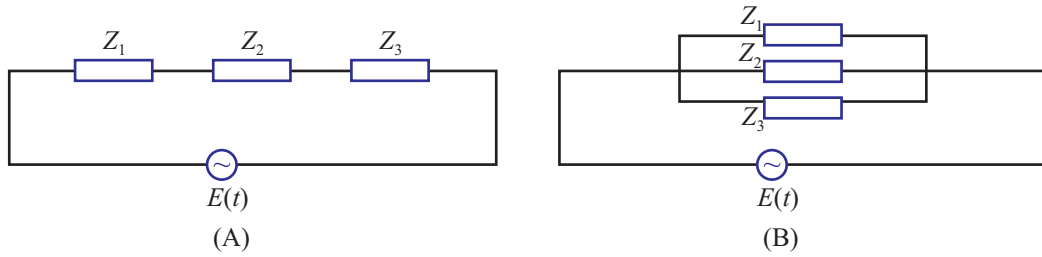


Figure 13-23: (A), (B): Series and parallel combination of three impedances Z_1 , Z_2 , and Z_3 , connected across an AC source $E(t)$; the symbols can be interpreted as complex quantities, with the tildes omitted.

By making repeated applications of the above two basic rules one can work out the equivalent impedance of a *network* of impedances in an AC circuit.

As a simple application, look at fig. 13-24 where there is a series L - R combination and a capacitor, the two being connected in parallel across an AC source. The series combination rule tells us that the L - R combination can be replaced with a single complex impedance $\tilde{Z}_1 = R + j\omega L$. Applying, then, the parallel combination rule to \tilde{Z}_1 and $\tilde{Z}_C = -\frac{j}{\omega C}$, one obtains the equivalent impedance of the L - C - R combination shown in the figure:

$$\tilde{Z}_{\text{eq}} = \frac{R + j\omega L}{1 + j\omega C(R + j\omega L)}. \quad (13-69)$$

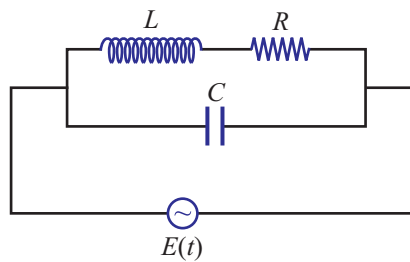


Figure 13-24: A circuit containing a parallel combination of a series L - R , and a capacitance C , connected across an AC source - illustrating the laws of series and parallel combination of impedances.

Problem 13-14

Check out the result (13-69). Work out the resistive and reactive parts of \tilde{Z}_{eq} .

Answer to Problem 13-14

HINT: Denoting the equivalent impedance of the series L - R combination by Z_1 and the equivalent impedance of the given network by Z_{eq} (we omit, for the sake of brevity, the tildes on the symbols representing complex quantities; moreover, as mentioned above, we use the symbol ' j ' instead of ' i ' to denote the square root of -1), one obtains by successive applications of the series and parallel combination formulas,

$$Z_1 = R + j\omega L, \quad \frac{1}{Z_{\text{eq}}} = \frac{1}{Z_1} + j\omega C,$$

from which follows (13-69). Thus, $Z_{\text{eq}} = \frac{R+j\omega L}{(1-\omega^2 LC)+j\omega CR}$. Multiplying the numerator and denominator with $(1-\omega^2 LC) - j\omega CR$, one obtains $Z_{\text{eq}} = R_{\text{eq}} + jX_{\text{eq}}$, where the equivalent resistive part is $R_{\text{eq}} = \frac{R}{(1-\omega^2 LC)^2 + \omega^2 C^2 R^2}$, and the equivalent reactive part is $X_{\text{eq}} = \frac{\omega L(1 - \frac{CR^2}{L} - \omega^2 LC)}{(1-\omega^2 LC)^2 + \omega^2 C^2 R^2}$.

NOTE: The reactive part of the equivalent impedance evaluates to zero when $\omega^2 = \frac{1}{LC}(1 - \frac{CR^2}{L})$. At times, the vanishing of the reactive part of the equivalent complex impedance of a network is taken to be the defining condition for *resonance*.

In analyzing an AC circuit one needs to know the AC current through and the AC voltage across each of the impedances in the circuit, where the two are related by eq. (13-66). For this one needs to employ the above-mentioned rules relating to the series and parallel combinations of impedances, along with a set of other rules such as the ones relating to voltage division and current division discussed, in the context of DC circuits, in section 12.6.2.

Kirchhoff's principles constitute a pair of useful rules for the analysis of AC circuits, as they do in the case of DC circuits (see section 12.7.1). An AC circuit is made up of resistors, inductors, capacitors, and AC sources of EMF, of which the first three are instances of *impedances*. The analysis of an AC circuit runs parallel to that of a DC circuit provided one replaces the resistances in the latter with complex impedances, DC EMF's with complex representations of AC EMFs, and DC currents with complex representations of AC currents. Thus, Kirchhoff's first and second principles for AC

circuits can be expressed in the forms

$$\sum \tilde{I} = 0, \quad (13-70a)$$

$$\sum \tilde{I} \tilde{Z} = \sum \tilde{E}, \quad (13-70b)$$

respectively, where the two equations refer to a junction and to a mesh in an AC circuit, and the tilde over a symbol indicates a complex representation.

Other convenient rules of a general nature, useful in the analysis of AC as also of DC circuits, are the *star-delta* transformation rules, and *Thevenin's* and *Norton's* principles. However, I will not enter into these rules and their applications in this introductory presentation of the basic principles underlying AC circuits.

Of basic relevance in the analysis of AC circuits is the *principle of superposition*, which is analogous to the same principle as applicable to a DC circuit, being a consequence of the fact that the instantaneous or the RMS value of the AC current through an impedance is proportional to the instantaneous or the RMS value of the AC voltage across it. The statement of this principle in the case of an AC circuit is entirely analogous to the corresponding statement for a DC circuit (sec. 12.7.2), with impedances replacing resistances, and AC currents and voltages replacing DC currents and voltages respectively.

Fig. 13-25 depicts an *AC Wheatstone bridge* made up four arms containing impedances $\tilde{Z}_1, \tilde{Z}_2, \tilde{Z}_3, \tilde{Z}_4$ (each of which can in turn involve more than one resistive and reactive elements; the figure depicts the equivalent impedance in each arm), the arrangement being similar to the DC Wheatstone bridge of sec. 12.7.3, with an AC source of EMF connected between the junctions A and C. The null condition (or the *balanced* condition) of the bridge corresponds to zero current through the AC ammeter A, which is achieved when the four complex impedances are related to one another as

$$\frac{\tilde{Z}_1}{\tilde{Z}_3} = \frac{\tilde{Z}_2}{\tilde{Z}_4}. \quad (13-71)$$

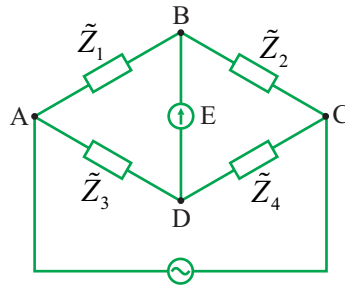


Figure 13-25: AC Wheatstone bridge; the arrangement is similar to the DC Wheatstone bridge shown in fig. 12-20; $\tilde{Z}_1, \tilde{Z}_2, \tilde{Z}_3, \tilde{Z}_4$ are the complex impedances in the four arms of the bridge; the balanced condition of the bridge (when the current through the AC ammeter E is zero) requires that the four impedances be related as in (13-71).

13.5.5 Power in an AC circuit

Equation (12-22) states that, in sending a current I through a circuit, the source of EMF supplies energy at the rate $\mathcal{E}I$.

This result was arrived at by considering a DC source of EMF \mathcal{E} and a steady current I . The result can be extended to a time-dependent EMF $E(t)$ and a time-dependent current $I(t)$ supplied by the source, by considering a small interval of time δt and making use of the above result to conclude that the energy supplied from the source in this small time interval is given by the expression $E(t)I(t)\delta t$. Summing up such expressions over a large number of such small successive time intervals, and dividing by the total time interval, one arrives at the *average* rate at which energy is to be supplied by the source in making possible the flow of current through the circuit.

The symbol \mathcal{E} is used to denote an EMF. However, as I have mentioned earlier, one also uses the symbol E for an EMF, where a careful reading of the context may be necessary at times to grasp which symbol is being used to denote which physical quantity. In this book you will find both the symbols having been used to denote EMFs.

For the sake of accuracy, one has to assume the time intervals δt to be vanishingly small,

in which case the average power supplied by the source reduces to the expression

$$\mathcal{P} = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \int_0^\tau E(t)I(t)dt, \quad (13-72)$$

where τ is the total time over which the averaging is performed. If the variation of the EMF and the current is a sinusoidal one (corresponding to the steady state in an AC circuit) then the above expression simplifies to

$$\mathcal{P} = \frac{1}{T} \int_0^T E(t)I(t)dt, \quad (13-73)$$

where T is the time-period, related to the angular frequency ω as $T = \frac{2\pi}{\omega}$. Recalling that there may, in general, be a phase difference between the applied EMF and the current, we substitute

$$E(t) = E_0 \cos \omega t, \quad I(t) = I_0 \cos(\omega t - \theta), \quad (13-74)$$

where E_0 and I_0 are the (real) amplitudes of the EMF and the current respectively. Substituting these in eq. (13-73) and working out the integral, one obtains

$$\mathcal{P} = \frac{E_0 I_0}{2} \cos \theta. \quad (13-75)$$

Notice that in deriving this result I did not make use of the complex representation of the EMF and the current. The complex representation is not of direct use here since the expression for the power involves a *product* of the EMF and the current, which is equivalent to a second degree expression in either of these two considered by itself.

An alternative expression for the power supplied by the source is

$$\mathcal{P} = E_{\text{RMS}} I_{\text{RMS}} \cos \theta, \quad (13-76)$$

where E_{RMS} and I_{RMS} stand for the *root mean squared values* of the EMF and the current. This relation is analogous to the relation $\mathcal{P} = EI$ (refer to eq. (12-22), where the notation is slightly different) for a circuit with a DC source of EMF E , carrying a DC current I ,

with the difference that eq. (13-76) involves the additional factor of $\cos \theta$, where θ stands for the phase lag of the current with reference to the EMF, i.e., equivalently, the phase angle of the complex impedance \tilde{Z} . Recall that the real impedance Z relates the real amplitudes I_0 and E_0 as $I_0 = \frac{E_0}{Z}$, which is obtained by taking the real parts of both sides of eq. (13-56a). Using the relation between the real amplitude and the RMS value of a sinusoidally varying AC quantity (refer to sec. 13.5.1.2), we can write the expression for the power in an AC circuit in the alternative forms

$$\mathcal{P} = \frac{E_{\text{RMS}}^2}{Z} \cos \theta = I_{\text{RMS}}^2 Z \cos \theta, \quad (13-77a)$$

or, making use of eq. (13-63),

$$\mathcal{P} = I_{\text{RMS}}^2 R, \quad (13-77b)$$

where R stands for the resistive part of the impedance \tilde{Z} . This means that the power supplied by the AC source is entirely accounted for by the heat dissipated in the resistive part of the AC circuit. However, if the circuit includes a component that converts electrical energy to some other form like, say, a motor then the above equation is to be modified into one of the form

$$\mathcal{P} = I_{\text{RMS}}^2 R + W, \quad (13-78)$$

where W stands for the average rate at which the energy supplied by the source is converted to forms other than heat dissipated in the resistive part.

The occurrence of the factor $\cos \theta$ in the expressions (13-77a) has led to the use of the name *power factor* for this quantity

Problem 13-15

For the AC circuit of fig. 13-24, assume that $R = 100\Omega$, $L = 5\text{mH}$, and $C = 2\mu\text{F}$. Obtain the value of the power factor. If the applied AC EMF is $E = E_0 \cos \omega t$, with the real amplitude $E_0 = 2\text{V}$ and angular frequency $\omega = 6000\text{s}^{-1}$, obtain the average power dissipated in the circuit.

Answer to Problem 13-15

HINT: Referring to problem 13-14 and substituting given values, one obtains the equivalent resistive and reactive parts of the complex impedance Z_{eq} (tilde omitted for the sake of brevity) as $R_{eq} = 54.1\Omega$ (approx.), $X_{eq} = -54.5\Omega$ (approx.), the latter telling us that the equivalent impedance is a *capacitive* one. The phase angle of the equivalent impedance is given by $\theta = -\arctan \frac{54.5}{54.1} = -\arctan 1.01$ (approx.), which gives the power factor $\cos \theta = 0.7$ (approx). The real expression for the current supplied by the AC source is $I = \frac{E_0}{|Z|} \cos(\omega t - \theta)$, which implies that the real amplitude of the current is $I_0 = \frac{E_0}{|Z_0|} = \frac{2}{\sqrt{54.1^2 + 54.5^2}} \text{A}$, i.e., 0.026A (approx). The average power dissipated is $P = \frac{I_0^2}{2} R_{eq}$ (refer to formula (13-77b), where the notation is different; recall that $I_{\text{RMS}} = \frac{I_0}{\sqrt{2}}$). This gives $P = \frac{0.026^2}{2} \times 54.1 \text{W}$, i.e., 18.3mW (approx).

13.5.6 The three-phase supply

Imagine that the set-up of fig. 13-6(B) is modified to include *three* independent coils instead of one, each consisting of conducting rods making up a rotating frame, and each with a pair of terminals for connection to an external load resistance. The resulting system of rotating coils is depicted schematically in fig. 13-26, where it is seen that the planes of the three frames are inclined at an angle of 120° with respect to one another.

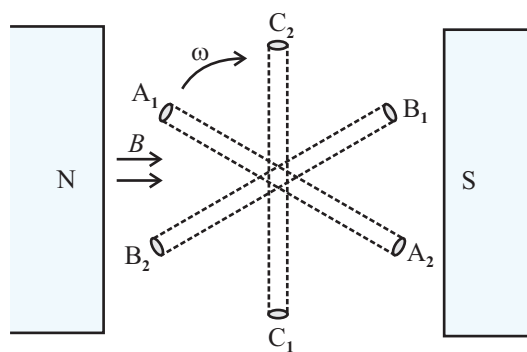


Figure 13-26: Depicting schematically the set of three rotating frames in a three-phase AC generator; $A_1, A_2, B_1, B_2, C_1, C_2$ denote cross-sections of the rods making up the rotating frames, where the planes of the three frames are inclined at an angle of 120° with respect to one another; the frames are made to rotate at the same angular velocity ω in a fixed magnetic field for which the direction of the lines of force is shown by the arrows.

Assuming for the sake of concreteness that the plane of the frame A_1A_2 is perpendicular to the direction of the magnetic field intensity at time $t = 0$ (this differs from the position shown in the figure), let the direction of rotation be such that the complex voltage (see sec. 13.5.1.3) between its terminals, denoted by a_1, a_2 respectively (see fig. 13-27 below), is given by

$$\tilde{V}_1 = V_0 e^{i\omega t}, \quad (13-79a)$$

the actual potential difference between the terminals being the real part of this complex-valued expression. Assuming that the terminals b_1, b_2 of the frame B_1B_2 are named in a manner analogous to the naming of the terminals a_1, a_2 of the frame A_1A_2 , the instantaneous complex voltage between these two terminals at time t is given by

$$\tilde{V}_2 = V_0 e^{i(\omega t + \frac{2\pi}{3})}, \quad (13-79b)$$

while, similarly, the complex voltage between the terminals c_1, c_2 of the frame C_1C_2 at time t is

$$\tilde{V}_3 = V_0 e^{i(\omega t + \frac{4\pi}{3})}. \quad (13-79c)$$

In other words, each of the three rotating frames acts as source of AC EMF for external loads that may be connected across their respective terminals. When connected to the external loads in an appropriate manner, the system of three rotating frames constitutes a *three phase* power supply. AC generators in power plants are mostly of the three phase type since a three phase power supply is endowed with a number of technical (and associated economic) advantages compared to a single phase supply where there is only one rotating frame in the magnetic field of the AC generator.

Problem 13-16

Show that the sum of the three potential differences V_1, V_2, V_3 of (13-79a)-(13-79c) at any given instant of time t is zero.

Answer to Problem 13-16

HINT:

$$V_1 + V_2 + V_3 = \operatorname{Re}(V_0 e^{i\omega t} (1 + e^{i\frac{2\pi}{3}} + e^{i\frac{4\pi}{3}})) = V_0 (\cos \omega t + \cos(\omega t + \frac{2\pi}{3}) + \cos(\omega t + \frac{4\pi}{3})) = 0.$$

Fig. 13-27 depicts schematically a three-phase generator where each coil (consisting of the rotating frame) is shown as a source of AC EMF (the internal resistance of the source, due to the resistance of the rods making up the frame, is not shown separately), and where the terminals a_2 , b_2 , c_2 are seen to be connected to a common point N. This mode of connection of the coils of a three phase AC generator is referred to as the *star connection*. In this case, of the lines leading from the coils to the external loads, one line is common so that, in total, there are *four* outgoing lines from the generator, where the line from the common terminal N is termed the *neutral* line. The other three lines are denoted as L_1 , L_2 , L_3 in fig. 13-27.

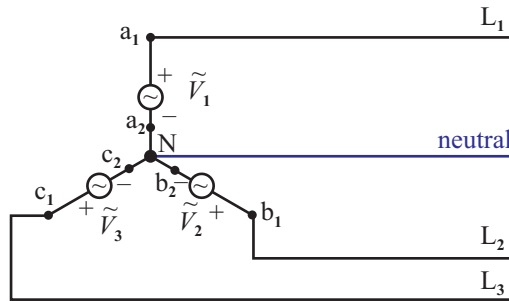


Figure 13-27: The three phase generator represented as three AC sources of EMF, each corresponding to one of the rotating coils; the star connection of the coils is shown where the three coils have a single common terminal N, referred to as the neutral terminal; four lines going out from the generator carry power to the loads, which may be located at distant points; the complex voltages across the three generator coils are \tilde{V}_1 , \tilde{V}_2 , \tilde{V}_3 , where the polarities for defining these voltages are shown; for instance \tilde{V}_1 is the difference between the complex voltage at terminal a_1 and that at a_2 , i.e., N.

Fig. 13-28 depicts the way the external loads are connected to the three sources of AC EMF provided by the three phase generator. The lines from a_1 and N connect to the load with complex impedance \tilde{Z}_1 between terminals a'_1 and N' , and two other loads

with impedances \tilde{Z}_2 and \tilde{Z}_3 are similarly connected between terminals b'_1 , N' , and c'_1 , N' . In this arrangement, the terminal N' , connected to N through the neutral line, is the common terminal of the three loads (the term 'load' is here used to denote the load impedances; more commonly, it is used to denote the load *currents*), and is termed the neutral terminal of the loads. Note that the loads are joined in a star type connection, corresponding to the star connection of the generator coils.

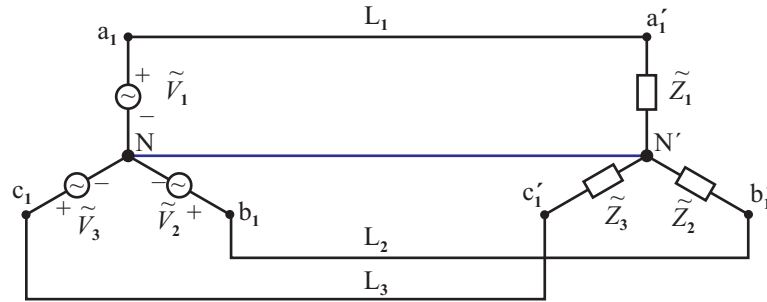


Figure 13-28: The loading of a three-phase generator, showing star connection of the generator coils as also of the load impedances; NN' is the neutral line, which carries zero current for a balanced loading; \tilde{Z}_1 , \tilde{Z}_2 , \tilde{Z}_3 are the three equivalent complex load impedances, each representing a number of loads in parallel.

The complex currents \tilde{I}_1 , \tilde{I}_2 , and \tilde{I}_3 in the loads are then given by

$$\tilde{I}_1 = \frac{\tilde{V}_1}{\tilde{Z}_1}, \quad \tilde{I}_2 = \frac{\tilde{V}_2}{\tilde{Z}_2}, \quad \tilde{I}_3 = \frac{\tilde{V}_3}{\tilde{Z}_3}. \quad (13-80)$$

If the complex impedances \tilde{Z}_1 , \tilde{Z}_2 , \tilde{Z}_3 connected as the loads are all equal (i.e., equal in their magnitude as also in the phase angles) then the loading of the three phase generator is said to be a *balanced* one. The actual loading of the generator may be unbalanced to some extent, causing technical problems with the operation of the generator.

With a balanced loading (i.e., with $\tilde{Z}_1 = \tilde{Z}_2 = \tilde{Z}_3 = \tilde{Z}$, say), the instantaneous current I in the neutral line turns out to be zero, since

$$\tilde{I} = I_1 + I_2 + I_3 = \text{Re}(\tilde{I}_1 + \tilde{I}_2 + \tilde{I}_3) = \text{Re}\left(\frac{V_0}{\tilde{Z}}(e^{i\omega t} + e^{i(\omega t + \frac{2\pi}{3})} + e^{i(\omega t + \frac{4\pi}{3})})\right) = 0. \quad (13-81)$$

The more the loading of the generator deviates from a balanced one, the larger will the

current in the neutral line NN' be.

Let us assume that the voltage drop across the three connection lines $a_1a'_1$, $b_1b'_1$, $c_1c'_1$ (denoted as, respectively, L_1 , L_2 , L_3 in fig. 13-28) are negligible, as are the internal drops across the generator coils. Then the voltages across the loads \tilde{Z}_1 , \tilde{Z}_2 , \tilde{Z}_3 , are the same as the voltages V_1 , V_2 , V_3 across the coils, which are the real parts of the corresponding complex voltages \tilde{V}_1 , \tilde{V}_2 , \tilde{V}_3 respectively. The amplitudes of each of these voltages being V_0 , the RMS value is given by $\frac{V_0}{\sqrt{2}}$. This is commonly referred to as the *phase voltage* (V_P) of the three phase power supply, and is the same as the RMS value of the voltage appearing across each of the loads.

The RMS value of the voltage between any two of the lines L_1 , L_2 , L_3 is referred to as the *line voltage* (V_L). For instance, the complex voltage between the lines L_1 and L_2 is $V_0(e^{i\omega t} - e^{i(\omega t + \frac{2\pi}{3})})$, and its amplitude is $\sqrt{3}V_0$ (check this out). In other words, the line voltage and the phase voltage are related as

$$V_L = \sqrt{3}V_P. \quad (13-82)$$

The same result is obtained if one considers, instead of L_1 , L_2 , the lines L_2 , L_3 , or L_3 and L_1 .

The RMS value of the current in any one of the lines L_1 , L_2 , L_3 is termed the line current (I_L), while that in any of the loads \tilde{Z}_1 , \tilde{Z}_2 , \tilde{Z}_3 is referred to as the phase current (I_P). For a balanced three phase supply in the star connection with load impedance \tilde{Z} in each of the lines, one has

$$I_L = I_P = \frac{V_P}{|Z|}. \quad (13-83)$$

Starting from the generating stations, the supply lines, including the neutral line, pass through sub-stations before reaching any particular locality where electrical power is to be supplied. To start with, the voltage produced by the generator is stepped up to a high value with a *transformer* (see sec. 13.5.7) and this high voltage is fed to the transmission line for being carried to distant locations, since the transmission of electrical power at a

high voltage has a number of advantages associated with it.

Briefly, if P is the power to be transmitted to a distant point, say, to a sub-station along any given line, then the current (I_L) in the line will be given by $P = V_P I_L$, i.e., transmission at a high voltage corresponds to a low line current. Since the line loss per unit distance covered is given by $P_{\text{loss}} = I_L^2 r$ (r being the line resistance per unit distance covered), a low value of I_L gives a correspondingly small power loss.

The voltage is then reduced by a process of step-down with the help of a transformer before the lines enter a locality for actual distribution to the consumers. Each of the three lines (along with the neutral line) carrying one of the three phases is connected to a number of establishments (houses, schools, offices), where it supplies to the loads of all these establishments connected in parallel across it. Each of these establishments, in turn, has a number of electrical appliances connected in parallel. All these loads connected in parallel give rise to an equivalent impedance, being one of the three impedances shown in fig. 13-28.

1. The AC generation and transmission of power is endowed with great flexibility and convenience because of the possibility of step-up and step-down that can be effected in the line voltage as required, where a transmission at a high voltage can be combined with generation and distribution at relatively low voltages.
2. The neutral line is earthed at the generating station. Each consuming establishment, in turn, is provided with an earth line (making a total of three lines for that establishment - the phase line, the neutral line, and the earth line). A potential difference between the neutral line and the earth line indicates that a current is flowing through the neutral line, in violation of the condition for a balanced power supply.

If I_{load} be the RMS value of the load current in any one of the appliances connected between the line and neutral of a given phase, then the power drawn by that appliance is given by $V_P I_{\text{load}} \cos \phi$, where $\cos \phi$ stands for the power factor of the appliance under consideration. Here the phase voltage V_P is the same for all the appliances in a given

phase since these are connected in parallel. On the other hand, the load currents and the power factors are different for the different appliances, and the total power supplied in a single phase is given by $V_P I_P \cos \theta$, where $\cos \theta$ is the effective power factor for the phase, given by $I_P \cos \theta = \sum I_{\text{load}} \cos \phi$, the summation being over all the appliances in the given phase.

Note that the sum of all the load currents I_{load} does *not* give the line current I_L , since the various different load currents may not be in the same phase.

The star connection is not the only way that the rotating coils in a three phase generator can be connected to distant loads. Another commonly used configuration is the *delta* connection, depicted in fig. 13-29 preferred for supply to establishments (such as factories) at a relatively high phase voltage. The delta connection makes use of three supply lines instead of four (the neutral line is absent) and economizes in the quantity of material required for the supply lines.

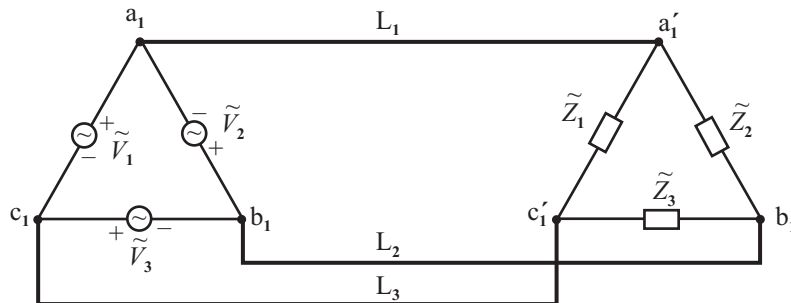


Figure 13-29: Delta connection; three lines (L_1 , L_2 , L_3) connect the terminals a_1 , b_1 , c_1 of the generator to the terminals a'_1 , b'_1 , c'_1 of the loads \tilde{Z}_1 , \tilde{Z}_2 , \tilde{Z}_3 .

The three phase supply is tailor-made for the production of rotating magnetic fields in synchronous and asynchronous AC motors (see sec. 13.2.4)

13.5.7 The transformer

The transformer is an AC machine of great practical value, making possible the stepping up and stepping down of AC voltages and efficient power transfer in electrical circuits.

The basic construction principle of a transformer is simple. It consists of two coils, or *windings* around a frame, with a *core* made of a ferromagnetic material like iron. The windings are insulated from one another and from the core. Of the two windings, one is connected to a source of AC voltage (and power) while the other delivers a stepped up or stepped down voltage to a load, with relatively small power loss. These are respectively the *primary* and *secondary* windings of the transformer, where the number of turns in these two windings, say N_1 and N_2 , determines the factor by which the primary voltage is stepped up or stepped down as it appears across the secondary.

Fig. 13-30(A) depicts schematically the two windings on the core, the AC source, and the load, while fig. 13-30(B) shows the circuit symbol of an iron-cored transformer. Air-cored transformers are used much less frequently.

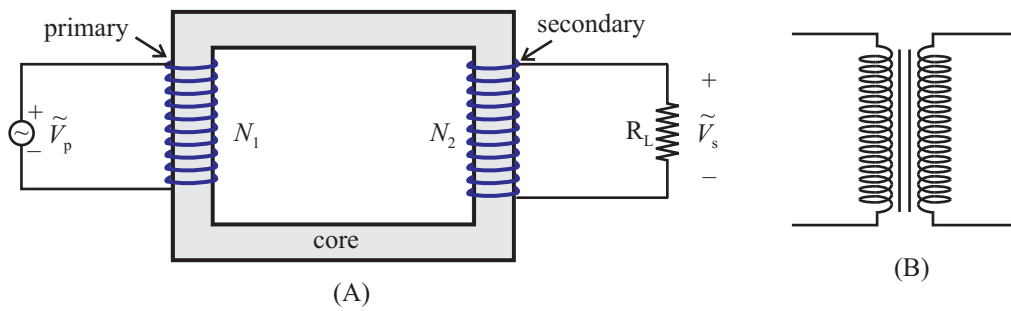


Figure 13-30: (A) transformer with load (schematic); the primary and secondary windings with N_1 and N_2 turns respectively, are wound around a core; the primary is fed from an AC source supplying a voltage \tilde{V}_p (in the complex representation), while the secondary effectively acts a source with voltage \tilde{V}_s for a load R_L (a resistive load is shown for the sake of concreteness); (B) circuit symbol of an iron-cored transformer.

13.5.7.1 Back EMFs

The resistances of the primary and secondary windings are mainly *inductive* in nature, and have small resistive parts that can be neglected in an approximate analysis of the working of the transformer. Let the EMF of the AC source in the primary circuit be $\tilde{V}_p = V_0 e^{i\omega t}$ in the complex representation, corresponding to $V_p = V_0 \cos \omega t$ in the real representation. The current in the primary winding sets up a magnetic field in the ferromagnetic core, corresponding to which the magnetic flux linked with each turn of

the primary winding is, say $\tilde{\Phi}$ in the complex representation. The flux linked with all the N_1 number of turns of the primary winding is then $N_1\tilde{\Phi}$. Since the flux changes with time with an angular frequency ω , there results an induced back EMF given by $-N_1 \frac{d\tilde{\Phi}}{dt}$. The time variation of the complex-valued quantities being of the form $e^{i\omega t}$, the time derivative is equivalent to a multiplicative factor $i\omega$, and the back EMF can be represented as $-i\omega N_1\tilde{\Phi}$.

The resultant EMF in the primary circuit is then $\tilde{V}_p + (-i\omega N_1\tilde{\Phi})$. Since the Ohmic drop across the primary winding is negligible (owing to the primary resistance being low), one has

$$\tilde{V}_p - i\omega N_1\tilde{\Phi} = 0. \quad (13-84)$$

Since the magnetic permeability of the core is very high, there occurs only a negligible loss of flux by leakage of the magnetic lines of force from the core, i.e., in other words, the flux linkage with the secondary winding is $\tilde{\Phi}$ per turn, implying a total flux linkage of $N_2\tilde{\Phi}$. Since this also changes with time with an angular frequency ω , the resulting back EMF in the secondary circuit is given by $(-i\omega N_2\tilde{\Phi})$. Recalling that the secondary circuit does not contain any external source of EMF, and that the Ohmic drop across the secondary winding can be neglected, the Ohmic drop across the load is given by

$$\tilde{I}_s R_L = -i\omega N_2\tilde{\Phi}, \quad (13-85)$$

where \tilde{I}_s is the secondary current in the complex representation, and where a resistive load R_L has been assumed for the sake of simplicity, so as to illustrate the basic principles involved.

subsubsectionThe voltage ratio

Thus, one arrives at the relation

$$\tilde{I}_s R_L = -\frac{N_2}{N_1} \tilde{V}_p. \quad (13-86)$$

In other words, the secondary winding effectively acts as a source of EMF with respect

to the load, given by

$$\tilde{V}_s = -\frac{N_2}{N_1}\tilde{V}_p. \quad (13-87)$$

This shows that the effective secondary EMF is in the opposite phase compared to the EMF of the external source in the primary circuit, and its magnitude is $\frac{N_2}{N_1}$ times the external EMF.

The phase of the voltage across the load can be reversed by reversing the connection of the load with the terminals of the secondary winding. Of the two possible choices of connection, one gives the above result, while the other gives a result with an added negative sign. What is of greater relevance here is the ratio of the magnitudes of the two voltages.

The effective EMF supplying to the load can thus be increased (stepped up) or decreased (stepped down) compared to the external EMF by appropriately choosing the turns ratio $\frac{N_2}{N_1}$.

13.5.7.2 The loading of the primary

Let the current in the primary circuit be \tilde{I}_1 in the complex representation. It might appear that, in spite of this current flowing in the primary, there is no net power delivered by the external source since the primary winding has been assumed to be an inductive one (implying that the power factor in the primary circuit is zero). In reality, however, there occurs a power loss in the load R_L of the secondary circuit, which has to come from the external AC source in the primary since there is no other source of power in the set-up. The power loss in the secondary is given by $\frac{V_s^2}{R_L}$, where V_s stands for the RMS value of the effective secondary voltage. Assuming ideal transformer operation with no power loss, this must be the power drawn from the AC source in the primary circuit. Making use of eq. (13-87), this can be written as

$$\mathcal{P} = \frac{V_p^2}{R_L \frac{N_1^2}{N_2^2}}. \quad (13-88)$$

In other words, the resistive load R_L in the secondary circuit of the transformer effectively introduces a resistive load

$$R_{\text{eff}} = R_L \left(\frac{N_1}{N_2} \right)^2, \quad (13-89)$$

in the primary.

Thus, the effective load seen by the external source of EMF in the primary circuit depends on the turns ratio $\frac{N_1}{N_2}$. This can be made use of in the *impedance matching* with respect to the internal impedance of the source. The term impedance matching refers to a requirement that has to be fulfilled for efficient transfer of energy from a source of EMF to a load presented to the source. For instance, if the source has a resistive internal impedance r (analogous to the internal resistance of an electrical cell), then it can transfer power efficiently to a resistive load R only if $R = r$, where a larger or smaller value of R causes a relatively larger power loss in the internal resistance of the source without that power being transferred to the load.

Since the effective load in the primary circuit of the transformer is $R_L \left(\frac{N_1}{N_2} \right)^2$, one can appropriately choose the turns ratio so that this matches with the internal resistance of the source of EMF in the primary circuit, as a result of which power is transferred efficiently from the source of EMF to the effective load in the primary, i.e., in the ultimate analysis, to the load R_L in the secondary. In other words, a transformer can be used for the purpose of impedance matching in a power transfer network.

13.5.7.3 The current ratio

The RMS value of the primary current is thus seen to be $I_p = \frac{V_p}{R_{\text{eff}}}$, and one finds, on making use of the relation (13-86),

$$\frac{I_s}{I_p} = \frac{N_1}{N_2}. \quad (13-90)$$

All these approximate results are borne out in a more detailed AC analysis of the primary and secondary circuits, provided a number of simplifying assumptions are made. More

precisely, if the secondary circuit contains a load impedance \tilde{Z}_L (which may have a reactive part in addition to the resistive part R_L), then an effective impedance $\tilde{Z}_{\text{eff}} = \tilde{Z}_L (\frac{N_1}{N_2})^2$ appears in the primary. The complex currents in the primary and the secondary are seen to be related by

$$\frac{\tilde{I}_s}{\tilde{I}_p} = -\frac{N_1}{N_2}, \quad (13-91)$$

while the complex voltages satisfy eq. (13-87).

The assumptions underlying the derivation of these relations are as follows: (i) no power loss occurs in the transformer; (ii) there is no flux leakage in the core, i.e., the mutual inductance M of the primary and the secondary coils is related to their self inductances (L_1, L_2) as $M = \sqrt{L_1 L_2}$ (refer to eq. (13-23)); (iii) the resistance (R_1) of the primary winding is negligibly small compared to ωL_1 , its reactance; (iv) the resistance (R_2) of the secondary winding is small compared to $|Z_L|$, which in turn is small compared to the reactance ωL_2 of the winding.

13.5.7.4 Energy losses in the transformer

Of the conditions listed above, the condition relating to the energy loss in the transformer needs special mention. Though, with present day technology, transformers are operated routinely at above 95% energy efficiency, even a small improvement in the efficiency will result in a considerable energy saving in absolute amount and will have a favorable environmental impact.

Energy losses in a transformer arise principally due to (a) Ohmic losses in the transformer windings, (b) magnetic hysteresis effects in the core, and (c) eddy currents.

The energy loss due to the heating of the wires used in the transformer windings can be minimized by using thick wires, though this results in a substantial increase in the cost. Magnetic hysteresis (see sec. 12.9.4.3) in the material of the core involves an energy loss since, in each cycle of the AC, the core goes through a cycle of magnetization and demagnetization which involves a loss of energy in the form of heat due to energy-

consuming irreversible processes involving the change in the magnetic domain structure within the material. The commonly adopted approach in the face of this problem is to make an appropriate choice of the material of the core, where the magnetization curve of the material is of the shape of a narrow loop, with a small value of the enclosed area since the area enclosed by the magnetization curve represents the hysteresis loss in one complete cycle of magnetization.

Amorphous steel is considered to be a useful material for the fabrication of the transformer core. This is a non-crystalline form of steel in which the hysteresis loss is minimized.

Eddy currents are circulating current loops set up within a conducting material as the magnetic field intensity in it is made to change with time (see sec. 13.5.8 below) where, from a basic point of view, the generation of these current loops is implied by Faraday's law of electromagnetic induction. The circulation of eddy currents within a conductor, in turn, implies an ' I^2R loss' of energy due to the Joule heating of the conductor. One way to reduce the eddy current loss is to laminate the material into thin slices where the slices are held together with an adhesive. This forces the eddy currents to circulate in smaller loops, causing a decrease in the heat loss. Transformer cores can even be designed to be made of iron *powder* where each granule in the powder is coated with an insulating material so that the eddy currents are confined within these granules.

13.5.7.5 The transformer in three phase distribution

In a three phase power distribution system, specially designed *three phase transformers* are commonly used, in which three sets of primary and secondary windings, each for one of the three phases, are wound on a common three-legged core as shown schematically in fig. 13-31. Though three separate transformers can be used in the place of a single three phase transformer, the latter is often more economical to use.

The three primary windings of a three phase transformer are fed from a three phase supply, while the three secondary windings, in turn, deliver power to a three phase

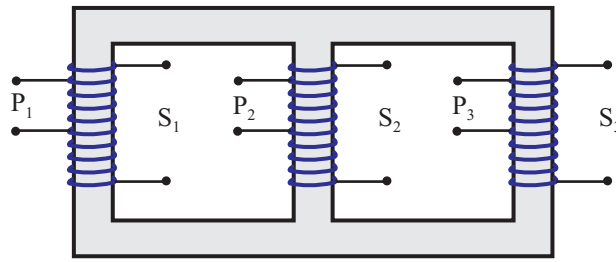


Figure 13-31: Three phase transformer, schematic; P_1 , P_2 , P_3 denote the three primaries, while S_1 , S_2 , S_3 are the corresponding secondaries wound on a single three-legged frame.

transmission system carrying power to distant loads. The three primary windings can be connected in a star configuration, being fed from four supply lines including a neutral. The three secondary windings, on the other hand, can be connected either in a star configuration feeding a four-line transmission system or in a delta configuration feeding a three-line supply system. Similarly, with the primary windings connected in a delta configuration to a three-line supply there are two possibilities for connecting the secondary windings. The transformer thereby offers a great flexibility in the designing of power grids, in addition to its impedance matching and stepping-up and stepping-down actions.

Fig. 13-32 shows a three phase transformer (which, as mentioned above, may be replaced with three single phase ones) connected in a star-star configuration, i.e., one in which the primaries as well as the secondaries are star-connected.

13.5.8 Eddy currents

Imagine a conductor placed in a changing magnetic field. Considering any closed path in the bulk of the conductor, the changing flux through the closed path implies an electromotive force around the path arising due to electromagnetic induction, i.e., a non-zero value for the line integral of the electric field intensity along the path. Since this holds for any and every closed path in the bulk of the conductor, one can equivalently say that an electric field is induced everywhere within the conductor and a resultant flow of current takes place, where the distribution of current is described by a current density

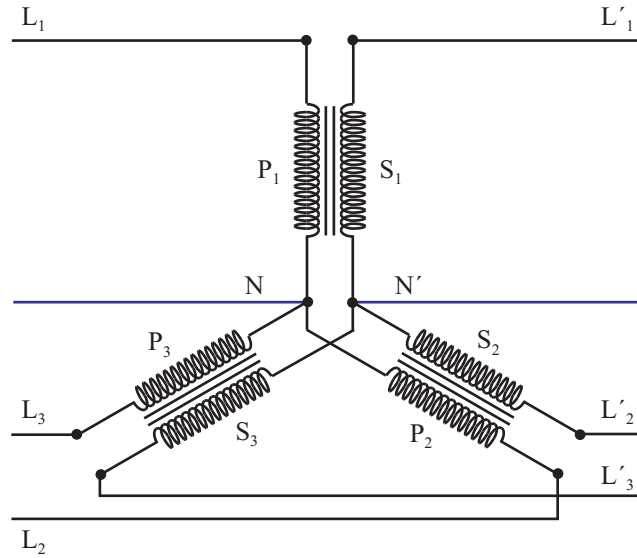


Figure 13-32: A three phase transformer (or three single phase transformers) in the star-star configuration; the primaries P_1 , P_2 , P_3 are star-connected, as are the corresponding secondaries S_1 , S_2 , S_3 ; the three lines L_1 , L_2 , L_3 along with the neutral line N feed the primaries while the secondaries, in turn, feed power into the lines L'_1 , L'_2 , L'_3 along with the neutral line N' .

\mathbf{j} that can vary from point to point.

Denoting the electric field intensity at any given point by \mathbf{E} and the conductivity at that point by σ , the current density is given by $\sigma\mathbf{E}$. Current loops are formed in the conductor representing the flow paths of current, where the loops may be of various shapes and sizes, which may, moreover, change with time. Such circulating currents in a conductor placed in a changing magnetic field are termed *eddy currents*.

Instead of a changing magnetic field in a stationary conductor, one may consider a conductor moving through a stationary magnetic field where essentially the same phenomenon of eddy currents is found to occur. The motional EMF over any and every closed path within the conductor once again leads to the generation of these eddy currents.

The description of eddy currents as circulating currents in discrete current loops is, however, of limited validity. A more general description is in terms of a *vector field* representing the spatial variation of the current density vector $\mathbf{j}(\mathbf{r})$ at every point \mathbf{r} within

the conductor at any given instant of time, where the vector field may be time dependent. A characteristic feature of this vector field is that the flux of $\mathbf{j}(\mathbf{r})$ over any closed surface is zero, from which the description in terms of closed current loops derives. However, instead of being discretely distributed in space, there may be a continuous distribution of such loops and, moreover, the geometry of the loops may be of a complex nature.

To gain an idea as to how the circulating current loops may be generated, let us consider a conducting disk being made to rotate through a magnetic field, the latter being confined to a rectangular area as in fig. 13-33, where the magnetic lines of force are perpendicular to the plane of the figure, directed into the plane from above. The motional EMF in the disc acts in the direction of $-\mathbf{v} \times \mathbf{B}$ (see sec. 13.2.2.2) which in the present instance corresponds to the direction from the bottom of the figure towards the top at the edges of the rectangular region occupied by the magnetic lines of force. Thus, circulating current lines are produced as shown by the arrows. Though closed current loops are shown in the figure, the current distribution in the conductor may not, in reality, be in the nature of a discrete collection of closed current loops.

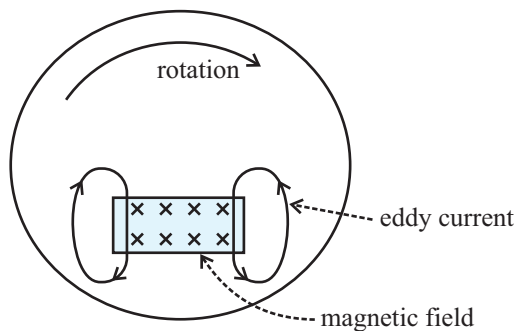


Figure 13-33: Illustrating the formation of eddy currents; a conducting disk made to rotate in a steady magnetic field; the field lines are confined to a rectangular region within the area of the disk, with the field lines perpendicular to the plane of the figure and directed inward; eddy current loops formed due to the motional EMF induced in the disk are indicated schematically, where a continuous distribution of such loops may be produced.

While the physics underlying the production of eddy currents is simple to describe, the geometry of eddy current loops may be of enormous complexity. There is, first of all, a *small scale* structure of the current distribution, governed by a large number of small

scale inhomogeneities in the medium, notably the crystalline imperfections in it. The distribution of these inhomogeneities and the resulting small scale structure of the eddy current distribution cannot be predicted by making use of any basic theoretical formula, and can only be described in *statistical* terms.

In describing the distribution of eddy currents on a *larger scale*, one has to average away much of the small scale structure, whereupon one finds that the vector field $\mathbf{j}(\mathbf{r})$ is determined by a set of partial differential equations relating to *Maxwell's equations* (see chapter 14) in the conducting material where local discontinuities in the material are no longer present. However, finding the solutions to these equations under appropriate boundary conditions is quite nontrivial. In particular the geometry of the vector field $\mathbf{j}(\mathbf{r})$ depends strongly on the shape and size of the conductor, especially on its *boundary surface*.

In this respect, the problem is analogous to the determination of the flow lines of a fluid enclosed in a container in which a convective motion is produced by means of heating or, to the determination of the excess pressure field for acoustic waves produced in a closed region with a boundary of arbitrary shape. The determination of the flow lines or the excess pressure field is a nontrivial problem in physics as is the determination of the large scale geometrical structure of the current lines in a conductor in a changing magnetic field. In particular, analogous to the turbulent flow of the fluid or the complex reverberation pattern in an auditorium, the spatial distribution of eddy currents may be of an apparently *random* nature which can once again be described only in statistical terms.

The generation of eddy currents in a conductors entails a *power loss* due to the heating up of the conducting material, which poses an engineering challenge in the construction of electrical machines such as transformers (see sec. 13.5.7.4). The commonly used solution to this problem is to laminate the conducting material into thin slices or sheets and to join up the sheets into a larger volume. Owing to the discontinuities in the material between the successive sheets, the eddy current loops break up into relatively smaller ones confined within the individual sheets. While this increases the number of

current loops, the energy loss within each loop is reduced by a larger factor, and the net result is a reduction of the total power loss.

On the other hand, eddy currents can be made use of in a number of home and industrial applications, a common application being in the induction heaters in home cooking.

Chapter 14

Wave motion II: Electromagnetic waves

14.1 Introduction

In chapter 9 I have given you a brief introduction to concepts relating to *waves* in physics. As a concrete example of waves, we looked at acoustic waves, where these can be described as pressure waves in a fluid.

The present chapter is a continuation of our study of waves where we are going to have a look at *electromagnetic waves*, these being waves associated with oscillations of electric and magnetic field strengths in a limited or extended region of space. Electromagnetic waves are of vast theoretical and practical importance in physics and are, moreover, of especial relevance since these provide the basis for the subject of *optics*.

While electric and magnetic fields appear to be distinct entities, where each can be described in terms independent of the other, at a deeper level the two are related to each other in an essential way. The descriptions of electric and magnetic fields that we have had so far are therefore parts of a more complete theory which tells us that, in reality, electric and magnetic fields affect and modify each other. It goes by the name of *electromagnetic theory* and its predictions have been found to be in conformity with

a large class of observed phenomena, leading to an enormous range of practical and technological applications.

In the present chapter I am going to present to you a few basic concepts in electromagnetic theory, and to employ these concepts in describing features of electromagnetic waves of relatively simple types. The next chapter will be devoted to an explanation of a number of optical phenomena on the basis of electromagnetic theory where the basic approach will be to make use of the fact that optical disturbances are made up of electromagnetic waves whose frequencies lie in a certain limited range. The resulting theory of optics is referred to as *wave optics*. The latter is a more fundamental theory of optical phenomena compared to the familiar *ray optics* outlined in chapter 10, and I will have a few words to say as to how ray optics relates to wave optics.

14.2 Electromagnetic theory

14.2.1 The electromagnetic field in free space

Electromagnetic theory starts from a description of electric and magnetic fields in vacuum in situations more general compared to the ones we have considered earlier in this book. The basis of this description is a set of four equations, collectively referred to as *Maxwell's equations*. Since the job I have set myself in this book is to present to you an elementary exposition, I will write down the equations for the sake of completeness, but we will not have occasion to make direct use of these. However, I mean to explain to you as far as I can what these equations stand for.

14.2.1.1 What the first equation means

Imagine a number of stationary charges located in vacuum in a certain region of space. These charges will then set up an *electric field* in vacuum. The electric field is completely described by specifying the electric field vector at all points in space which is achieved by making use of the basic formula (11-6a). If, instead of a number of discrete point sources, the electric field is produced by a continuous distribution of charges, then that distribution can be imagined to be made up of a large number of small volume elements,

each acting effectively as a point charge, and the electric field intensity at any given field point can once again be worked out along similar lines, making use of Coulomb's law and the superposition principle.

As pointed out in section 11.8, an alternative description of the electric field is obtained from Gauss' principle which, in many respects, provides a more convenient description of the electric field. On the face of it, it looks more complicated than Coulomb's law since, instead of straightaway giving you an expression for the electric field intensity in terms of the source charges, it relates the surface integral of the intensity over the closed boundary of a region to the charges residing inside that region. Even so, Gauss' principle, or an alternative form of it, known as the *differential form* of Gauss' principle, is commonly chosen as one of the basic principles of electromagnetic theory.

However, even though the differential form of Gauss' principle appears to be just a re-statement of the basic principles of electrostatics that we had a look at in chapter 11, the *context* in which it is used in electromagnetic theory is quite different. While in electrostatics, all charges are to be considered as stationary, electromagnetic theory allows for *moving charge distributions*. Indeed, electromagnetic theory allows for charges *and* currents as sources where *both* the charge and current distributions may *vary with time*. With this big difference in the general setting, the differential form of Gauss' principle looks *formally* the same in electromagnetic theory as in electrostatics. This fact, that the differential form of Gauss' principle survives the altered setting which now admits of time-varying charge densities, currents, and field strengths, is indeed of notable significance. It constitutes the *first* basic equation in electromagnetic theory.

The mathematical expression of the first equation, for an electromagnetic field set up in free space, is of the form

$$\text{div } \mathbf{E} = \frac{\rho}{\epsilon_0}, \quad (14-1)$$

where ρ stands for the charge density at any given point at any specified instant of time, and \mathbf{E} for the electric field intensity, the symbol 'div' denoting the divergence operator

(refer to section 2.14.1 for the definition of divergence of a vector field) at the space-time point under consideration.

14.2.1.2 What the second equation means

The second basic principle in electromagnetic theory is once again a familiar principle, one from the theory of stationary magnetic fields. Recall that a stationary magnetic field is produced by *steady currents*. Since steady currents flow in closed loops, like the current in a closed loop of wire set up by a Galvanic cell, a basic feature of stationary magnetic fields is that the magnetic lines of force are *closed lines*. An alternative expression of the same feature is that there is no analogue in magnetism of the electric charge, referred to as a *monopole*, acting as the source of the magnetic field. If you try to construct an analogue of Gauss' principle in magnetism, then you will end up with the result that the surface integral of the magnetic field intensity (\mathbf{B}) over a closed surface is always *zero*. This feature of magnetic fields persists even when the distribution of source currents varies with time, and it can once again be given an alternative, *differential* form. This, then, constitutes the *second* basic equation in electromagnetic theory, whose mathematical form is

$$\text{div } \mathbf{B} = 0. \quad (14-2)$$

14.2.1.3 What the third equation means

The third basic equation in electromagnetic theory is based on *Faraday's law of electromagnetic induction*. The familiar form of this principle states, for instance, that an electromotive force is set up in a closed loop of wire if the magnetic flux through it changes with time. Now, a source of EMF is a curious thing in that it implies, in some way or other, a breakdown of the feature that the electric field is *conservative* and can be expressed in terms of a *potential*. This last feature finds expression in the mathematical statement that the line integral of the electric field vector over a closed line is zero. However, this line integral evaluated over a closed circuit containing a source of EMF relates precisely to the EMF in the circuit and hence, for a circuit with a source of EMF in it, the line integral *cannot* be zero.

In the case of a Galvanic cell acting as the source of EMF setting up a current in a wire connected between its terminals, the description of the electric field in terms of an electrical potential breaks down at the electric double layers surrounding the two electrodes, since the potential assumes two different values as a double layer is approached from two sides. A more complete description involves the *electrochemical* potential which, in contrast to the electrical potential, has a unique value everywhere in the circuit.

In a situation involving time-varying magnetic fields, the line integral of the electric field intensity turns out to be non-zero not only along a closed loop of wire or a closed circuit but, in general, over *any* closed path. The implication is that the description of the electric field in terms of the potential is no longer possible where, more specifically, the line integral of the electric field intensity over any closed path is related to the rate of change of magnetic flux associated with any surface with that closed path as its boundary. Once again, this principle finds an alternative expression in a *differential* form which, then, constitutes the *third* basic equation in electromagnetic theory, expressed as

$$\text{curl } \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \quad (14-3)$$

where the curl operator has been introduced in section 2.14.1, and the right hand side of the equation involves the partial derivative of $\mathbf{B}(\mathbf{r}, t)$ with respect to t , with the position vector held fixed.

14.2.1.4 What the fourth equation means

The fourth basic principle constituting the foundations of electromagnetic theory is a generalization of the statement that currents are the sources of magnetic fields where, by the term current, one means charges in motion. However, consider the circuit shown in fig. 13-18 where a Galvanic cell is connected in a circuit made up of a resistor (say, a piece of conducting wire) and a *capacitor*, say, one made of two parallel plates. When the key K in the circuit is closed, there appears a transient phase when charges pile up on the plates of the capacitor, causing the potential difference between the plates to

increase. Since the plates of the capacitor are insulated from one another (the intervening medium may even be vacuum) no charges move between the plates and hence no current, in the familiar sense, flows between them. Yet, if a magnetic needle is held between the plates, it will be found to register a deflection, indicating that a magnetic field has been set up.

This is where Maxwell generalized the idea of currents, stating that, in addition to the familiar *conduction currents* resulting from the motion of charges, one has to consider *displacement currents* that can also act as sources of magnetic fields, where he related displacement currents to *changing electric fields*. For instance, in the above example of the capacitor plates, a varying potential difference between the plates means a changing electric field strength, which results in a displacement current being set up between the plates and therefore in a magnetic field. This generalization of Maxwell's shows up as a modification of *Ampere's circuital law* (see section 12.8.10). Recall that the latter can be made use of in describing a magnetic field in terms of the current distribution acting as its source, being an alternative principle compared to the one presented in sec. 12.8.6. Indeed, Ampere's circuital law plays a similar role in magnetism as does Gauss' principle in electrostatics.

Thus by introducing the notion of the displacement current, Maxwell made possible a generalization of Ampere's circuital law, and thereby a description of magnetic fields in situations involving changing electric fields. This constitutes the fourth basic equation in electromagnetic theory where, once again, a differential form of the modified Ampere's law is commonly used. The mathematical form of the fourth basic equation of electromagnetic theory, which is the differential form of the circuital law, with the displacement current taken into account, is

$$\text{curl } \mathbf{B} = \mu_0 \left(\mathbf{j} + \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right), \quad (14-4)$$

where the second term within the brackets on the right hand side stands for the displacement current.

14.2.1.5 The four equations: an overview

While I have refrained from examining the four basic equations of electromagnetic theory in details, I can still give you an idea as to what use these equations are put to.

The equations tell us how the space- and time derivatives of the field vectors are related among themselves and, moreover, imply that a changing electric field has a magnetic effect while, similarly, a changing magnetic field has an electrical effect. The two fields thereby cease to be independent entities and appear as *coupled* to each other. In other words, the electric and magnetic fields, instead of being distinct and independent of each other, now appear as manifestations of a single entity, the *electromagnetic field*.

Looked at from this point of view, Maxwell's equations are a set of equations governing this single composite entity as a *dynamical system*, in much the same way as Newton's equations of motion govern the dynamics of a particle or a system of particles. Recall that Newton's equations are a set of differential equations that one has to solve, making use of appropriate initial conditions, to know how the state of a given dynamical system evolves in time. The basic difference between a system of particles and an electromagnetic field as a dynamical system is that, while a particle or a set of particles requires only a relatively small number of variables for the description of its state, the state of the electromagnetic field requires the specification of electric and magnetic field intensity vectors at *all* points in space, even ones at *infinitely* remote regions.

This makes the Maxwell equations a somewhat more complex set of equations but there is not much of a difference in substance - one has to solve the Maxwell equations, making use of appropriate initial *and* boundary conditions to get to know how the field variables change in time *and* space.

As an example, consider the radiation of electromagnetic waves from an *antenna* in which an alternating current is set up with the help of an AC source of EMF. Depending on how the current in the antenna is set up, one can write out Maxwell's equations in the space surrounding the antenna, treating it as vacuum for the sake of simplicity, the antenna being represented as a *source* term in these equations, in much the same

way as stationary charges are treated as sources for working out the strength of an electrostatic field at any given point in space. One can then solve these equations subject to appropriate *boundary* conditions - conditions that represent, in this instance, the expected behavior of the field at large distances from the antenna.

What results from this exercise is the *general solution* under these boundary conditions. This can be made use of to arrive at a *particular solution*, giving the values of the field variables at all points in space and at all possible instants of time, when the appropriate *initial conditions*, viz., the values of the field variables at any given initial point of time, are known (alternatively, a solution can be determined by assuming its time variation to be a harmonic one, with a given frequency).

In summary, then, Maxwell's equations are a set of differential equations governing the behavior of electromagnetic fields. These equations can be solved by following appropriate mathematical procedures so as to work out how the electric and magnetic field intensities vary in space and time in any given situation involving given sources, where the latter include time dependent charge- and current distributions in space. One commonly emerging feature of the solutions to Maxwell's equations thus arrived at is that variations of the field intensities initiated in any one region of space are *transmitted* to other regions in the form of *waves*.

In the remaining part of the present chapter I will introduce simple wave solutions of Maxwell's equations; but first I have to tell you a bit about electromagnetic fields in material media.

14.2.2 Electromagnetic fields in material media

While electromagnetic fields in vacuum are comparatively simple to describe, the fields are more commonly set up in some material medium or other. The space-time behavior of the fields, including the interdependence of the electric and magnetic field components, in a material medium is described once again by a set of equations which are generalizations of the four basic equations for free space, and are referred to as Maxwell's equations in the medium. While these equations look similar for various dif-

ferent media, in reality, the equations for any given medium involve features specific to that medium. In other words, while the electromagnetic fields in the different media have some features in common, each medium leaves its own imprint on the space-time behavior of the fields.

For instance, as we will see below, *plane* waves can propagate in any given medium and have features similar to plane waves set up in vacuum, but the speeds differ from one medium to another and, moreover, the speed of a monochromatic wave in a medium *depends on the frequency*. Another feature distinguishing an electromagnetic field in a material medium from that in vacuum is that there occurs *absorption* in a medium whereby the energy associated with the field gets attenuated and an equivalent amount of energy appears in the medium in the form of heat. In a *conducting medium* for instance, a plane wave is quickly damped out, with the intensity decreasing as the wave propagates through the medium.

14.3 Electromagnetic waves

All matter is made up of *charged particles* in continual motion. These microscopic charges, and the microscopic currents generated by these charges, create an electric and magnetic field everywhere in space. These electric and magnetic fields are *coupled* to each other in that a variation of the magnetic field intensity generates an electric field, and conversely, a variation of the electric field intensity generates a magnetic field. Taken together, the electric and magnetic fields in space constitute a single dynamical system, the *electromagnetic field*. As I have already pointed out, the behavior of this dynamical system is completely described by the Maxwell equations for the medium under consideration.

The electromagnetic field in space is never in a stationary condition, but in a state of continual change or *disturbance*, much like the water surface of an ocean, where a vast multitude of disturbances keep on occurring at various places due to innumerable local causes. Once a disturbance is created, it *propagates* in space in the form an *electromagnetic wave*, in conformity with Maxwell's equations describing the situation

under consideration. In general, the space-time dependence of solutions of Maxwell's equations may be quite complex. At times, the term 'wave' is employed to denote a class of relatively simple solutions having a number of common features, while the more general and complex types of solutions that can be described only in statistical terms, are referred to as electromagnetic disturbances, or by a term to a similar effect.

14.3.1 Sources of electromagnetic waves

What can be the local cause creating a disturbance in the electromagnetic field at any particular location in space? The most common and general local cause corresponds to *changes in the states of motion* of electrons in material bodies located in some regions of space or other.

The electrons in all materials, including electrons in the atmosphere or even in a chamber with a high degree of vacuum in it, are in continuous motion. These electrons can be classified into two groups, namely, *bound* and *free* electrons. While the former are attached to particular atoms or groups of atoms, the latter can move away to large distances by themselves.

According to the quantum theory, a bound electron can be in any one of a number of stationary states (see chapter 16) with *discrete* energy values. Such an electron can occasionally make a *transition* from a higher energy state to a lower energy one due to the influence of other atoms, molecules or *photons*. Such transitions cause a disturbance in the electric and magnetic fields in space around the atom or group of atoms to which the electron is bound, which then propagates in the form of an electromagnetic wave.

Free electrons also suffer changes of their states of motion when they are accelerated or decelerated by, say, an applied electric field, or when they are *scattered* in collisions with other particles. Such changes also generate electromagnetic disturbances that propagate in the form of waves. For instance, an alternating current set up in a transmitting antenna causes the free electrons in the antenna to continuously undergo changes in their states of motion, which leads to electromagnetic waves being set up and energy being carried away from the antenna.

A material body that causes the generation of a disturbance in the electromagnetic field in a region of space around it by virtue of changes in states of motion of microscopic charges in it, is referred to as a *source* of electromagnetic radiation.

In electromagnetic theory, a source is described in terms of the variation of *charge density* and *current density* in some region of space, these charge- and current densities being defined from a *macroscopic* point of view. From a *microscopic* point of view, on the other hand, the source involves changes in the states of motion of charged constituents of matter in the given region of space. In general, these innumerable changes cannot be described by charge and current density variations in a simple way. One then describes the source in *statistical* terms, specifying a number of statistical averages of the charge and current density fluctuations.

14.3.2 Transmission of energy

The setting up of electric and magnetic fields in space require the expenditure of *energy*. Looked at from the point of view of energy, the generation and propagation of electromagnetic disturbances is explained as follows. The changes in the states of motion of the electrons (or other microscopic charges) in a source generating the electromagnetic waves, are generally accompanied by a decrease in the energies of these electrons, whereby some energy is made available for the setting up of electromagnetic disturbances around the source. Parts of this energy then flow to various distant regions of space by means of electromagnetic waves. Indeed, this *transmission* of energy from one region of space to other distant regions is a main characteristic of electromagnetic (as also of other) waves.

In some situations, however, an electromagnetic disturbance does not involve a transmission of energy through space. These are referred to as *stationary waves*.

14.3.3 The principle of superposition

In certain special situations, Maxwell's equations are found to possess wave solutions of a comparatively *simple* nature. Examples of such simple wave solutions that happen

to be of considerable practical relevance, are the *monochromatic plane wave* and the *monochromatic spherical and cylindrical waves*, where the term ‘monochromatic’ means that the oscillations of the electric and magnetic field intensities at any given point in space occur with some particular single frequency.

A large class of waves of a more complex nature can be expressed mathematically as a *superposition* of the relatively simple monochromatic plane, spherical, or cylindrical waves. Indeed, Maxwell’s equations possess the property that, knowing two solutions of these equations, one can construct a third solution by taking the sum, or superposition, of these two solutions. Thus, starting from a number of monochromatic plane wave solutions, one can obtain a solution of a more general nature by their superposition. The resulting wave may involve a number of frequencies rather than one single frequency, and is termed a *non-monochromatic* one.

The nature of the wave generated in a given set-up depends on the source, the medium, and the presence of other material bodies, and in most situations an exact solution of Maxwell’s equations describing the wave cannot be obtained. One then works out *approximate* solutions involving superpositions of relatively simple wave solutions. For a large number of situations, especially those involving waves of small wavelengths (e.g., in optics) a representation in terms of superpositions of relatively simple wave solutions is not at all possible, since the electromagnetic field is generated by a large number of microscopic sources generating electromagnetic disturbances in an uncorrelated manner. What is then possible is only a statistical description of the electromagnetic disturbance. In such situations a term like ‘electromagnetic disturbance’ is more commonly used in place of ‘electromagnetic wave’.

In the following, I will concentrate on the simplest of wave solutions of Maxwell’s equations, namely the plane and spherical monochromatic waves. Electromagnetic fields that can be approximately described by such solutions can be generated by making use of special set-ups and devices. Since more complex solutions to Maxwell’s equations can be made up by making use of these waves as building blocks, a study of these simple solutions is of great theoretical importance in understanding more complex situations

in electromagnetic theory and in optics.

From a mathematical point of view, a ‘wave’ is a function that satisfies a differential equation of a certain type, termed a ‘wave equation’. In electromagnetic theory, a wave equation is seen to arise as a corollary of Maxwell’s equations. In other words, solutions of Maxwell’s equations do satisfy a wave equation, and hence qualify as waves in the mathematical sense of the term.

14.4 The plane progressive monochromatic wave

The plane progressive monochromatic wave is one of the simplest solutions of Maxwell’s equations describing the variation of electric and magnetic field intensities with time at all points in space, either in vacuum or, more generally, in a homogeneous material medium.

14.4.1 Space-time field variations

Consider, for instance, the following expression for the electric field intensity at any point with position vector \mathbf{r} at a time instant t :

$$\mathbf{E}(\mathbf{r}, t) = \hat{e}_y E_0 \cos(kx - \omega t). \quad (14-5a)$$

This equation resembles the formula (9-4) describing a plane monochromatic acoustic wave in a medium with the difference that, while for an acoustic wave the wave function (i.e., the physical quantity whose oscillations are transmitted from one region to another in space) is a scalar, the electric field intensity which constitutes the wave function in the present instance is a *vector*. As the above equation tells us, the electric field intensity is, at all points in space and all instants in time, directed along the y-axis (\hat{e}_y being the unit vector along the y-direction of any chosen right handed Cartesian co-ordinate system), i.e., the Cartesian components E_x and E_z are both zero. One can then re-write

the equation in the simpler form

$$E_y = E_0 \cos(kx - \omega t) \quad (E_x = E_z = 0). \quad (14-5b)$$

which resembles eq. 9-4 more closely, and where the arguments (\mathbf{r}, t ; or, x, y, z, t) in E_y (or in E_x, E_z) have been omitted for the sake of simplicity.

Of course, an electromagnetic wave is not described completely just by specifying the variation of the electric field intensity in space and time. What one needs in addition is a description of the *magnetic* field intensity as well. Consider, then, the following equation describing the magnetic field intensity,

$$B_z = B_0 \cos(kx - \omega t) \quad (B_x = B_y = 0). \quad (14-5c)$$

The description of the wave is completed by appending the following information

$$B_0 = \frac{E_0}{v}, \quad \frac{\omega}{k} = v, \quad v = \frac{c}{n}, \quad (14-5d)$$

where v and n are appropriate constants depending on the medium under consideration and c is a universal constant referred to as the *velocity of light in vacuum*. The electric and magnetic fields described by the above equations are seen to satisfy Maxwell's equations for the medium. The field variations here describe a *plane progressive monochromatic wave* propagating along the x -axis of the Cartesian co-ordinate system under consideration.

It is supposed that the medium under consideration, in addition to being a homogeneous one, is also *isotropic*, i.e., its properties are the same in all possible directions in the medium. Crystalline materials and certain biological samples, made up of *macromolecules* provide instances of *anisotropic* media where the propagation of plane electromagnetic waves is characterized by a number of special features. Additionally, we assume that the medium under consideration is *non-dispersive*, and that the extent of *absorption* in the medium is negligible (see sec. 14.7 for basic ideas relating to dis-

persion and absorption, where a number of features of propagation of plane waves in dispersive media are outlined).

An isotropic homogeneous medium is characterized by two constants, namely, its *relative permittivity* (ϵ_r) and its *relative permeability* (μ_r) depending on its material properties. The constant v is related to these as

$$v = \frac{1}{(\epsilon_r \mu_r \epsilon_0 \mu_0)^{\frac{1}{2}}}, \quad (14-6)$$

where ϵ_0 and μ_0 , introduced earlier, are two universal constants, referred to as the permittivity of free space and the permeability of free space respectively (see chapters 11 and 12). In the case of vacuum, one has $\epsilon_r = 1$, $\mu_r = 1$, and then v reduces to $\frac{1}{(\epsilon_0 \mu_0)^{\frac{1}{2}}}$, which is nothing but the velocity of light (and of all electromagnetic disturbance) in vacuum.

The relations (14-5b) and (14-5c) tell us that the unit vectors directed along *the electric field intensity, the magnetic field intensity, and the direction of propagation form a right handed orthonormal triad*. Here the electric and magnetic field intensities may be taken as the instantaneous vectors or their vector amplitudes ($\hat{e}_y E_0$, $\hat{e}_z B_0$) with appropriate phase factors (i.e., in the present instance, appropriate signs of E_0 and B_0 , where both are positive for the wave under consideration; other choices for the signs of E_0 and B_0 also correspond to monochromatic plane waves).

Problem 14-1

The instantaneous magnetic field intensity components for a plane electromagnetic wave in vacuum propagating along the x-axis are $B_x = 0$, $B_y = 0.5 \times 10^{-8} \cos(2\pi(10^{15}t - ax))$, $B_z = 0$, where all quantities are in SI units. Find the value of a and the components of the instantaneous electric field vector. Calculate the time average of the product $\frac{1}{\mu_0} |\mathbf{E} \times \mathbf{B}|$.

Answer to Problem 14-1

Here the angular frequency is $\omega = 2\pi \times 10^{15} \text{ s}^{-1}$, and hence $a = \frac{k}{2\pi} = \frac{\omega}{2\pi c} = \frac{1}{3} \times 10^7 \text{ m}^{-1}$ (thus the

wavelength is $\lambda = 3 \times 10^{-7}$ m, see sec. 14.4.2). Since the wave is propagating along the positive direction of the x-axis, and since the electric field intensity, the magnetic field intensity, and the direction of propagation form a right handed triad, the electric intensity vector points along the negative direction of the z-axis. Further, the magnitudes of the electric and magnetic vectors are related as $E = cB$. Thus, the instantaneous electric field components are $E_x = 0$, $E_y = 0$, $E_z = -1.5 \times \cos(2\pi(10^{15}t - ax))$, with a obtained above. The vector $\mathbf{E} \times \mathbf{B}$ is directed along the x-axis and its instantaneous magnitude is $\frac{(1.5)^2}{3 \times 10^8} \cos^2(2\pi(10^{15}t - ax))$. Since the time average of the squared cosine term is $\frac{1}{2}$, the average value of the magnitude of the expression $\frac{1}{\mu_0} \mathbf{E} \times \mathbf{B}$ is 2.98×10^{-3} W·m⁻² (approx) (this is the time averaged Poynting vector at any given point (refer to eq. (14-10) below), and gives the average rate of energy transport per unit area, i.e., the intensity of the plane wave, at the point).

14.4.2 Frequency, wavelength and velocity

I will now turn to the explanation and interpretation of equations (14-5b)-(14-5d) describing a plane monochromatic electromagnetic wave, where you will find a close analogy to the plane monochromatic acoustic wave described in chapter 9.

Recall, incidentally, our use of the term ‘wave function’ in chapter 9. I have made use of this term to denote any physical quantity whose variation in space and time corresponds to a wave, i.e., an oscillation in time that is transmitted through space. More specifically, the quantity is required to satisfy a *wave equation*, the latter being a differential equation of a certain type. The term is, however, commonly used in a more restricted sense, to denote a function describing the state of a *quantum mechanical* system. The wave function in quantum theory satisfies a wave equation analogous to but differing from the equation satisfied by the excess pressure for an acoustic wave or the electric and magnetic field intensities for an electromagnetic wave. I will introduce the quantum theoretic wave function later, in chapter 16. For now, the term ‘wave function’ will be used in the broader sense of a physical quantity that satisfies a wave equation. In electromagnetic theory, the wave equation arises as a corollary of Maxwell’s equations.

Note first of all in the above expressions that the electric and the magnetic field intensities at any point \mathbf{r} with co-ordinates x , y and z , while remaining directed along the y - and the z -axes respectively, *oscillate* sinusoidally in time with a time-period $T = \frac{2\pi}{\omega}$ (frequency $\nu = \frac{1}{T} = \frac{\omega}{2\pi}$). The electric field intensity E_y along the y -axis oscillates between E_0 and $-E_0$, where E_0 (assumed to be positive for the sake of concreteness) is termed the *amplitude* of the electric field strength. Similarly, B_z oscillates between B_0 and $-B_0$ where B_0 represents the amplitude of the magnetic field intensity. The wave solution represented by the above equations is termed *monochromatic* because it is characterized by a specific value (ν) of the frequency.

On the other hand, at any given instant of time, the variation of E_y or B_z with x is also a sinusoidal one. The value of the wave function (E_y or B_z) at any given point is repeated after intervals of length $\lambda = \frac{2\pi}{k}$ along the x -axis, where λ is termed the *wavelength* of the plane monochromatic wave under consideration. Plotting the wave function against x for a particular value of t gives a sinusoidal curve as in fig. 9-3, which we term the 'wave profile' at time t . If we plot the wave profile at some other time, say, $t + \tau$, it will be seen to have an identical form, but shifted by a certain distance along the x -axis. The amount of shift is seen to be of the form $\xi = v\tau$, where v depends, in addition to the physical properties of the medium, on the frequency (ν) of the electromagnetic wave as well. It is termed the *phase velocity* of the wave. I will come back to the phase velocity in a while, after I introduce the concept of the *phase* of the wave. In the case of a wave propagating in vacuum, the phase velocity v reduces to c , the *velocity of light in vacuum*.

Since the wave profile propagates along the x -axis (with the phase velocity v), it is termed a progressive wave. Denoting the unit vectors along the three Cartesian axes by \hat{e}_x , \hat{e}_y and \hat{e}_z , note that the electric and magnetic field intensity vectors \mathbf{E} and \mathbf{B} are along \hat{e}_y and \hat{e}_z (or $-\hat{e}_y$ and $-\hat{e}_z$, depending on the sign of E_0 and B_0) respectively while the direction of propagation of the wave profile is along \hat{e}_x . An alternative way of relating the direction of propagation with \mathbf{E} and \mathbf{B} is to say that the wave propagates in the direction of the vector $\mathbf{E} \times \mathbf{B}$. In other words, as mentioned earlier, \mathbf{E} , \mathbf{B} , and the unit vector along the direction of propagation form a *right-handed orthogonal triad* of vectors. The fact that the electric and magnetic field intensity vectors oscillate in a plane perpendicular to the

direction of propagation, is expressed by saying that the wave under consideration is a *transverse* one.

We will presently see that the direction of propagation of the wave profile can be alternatively described as the direction of propagation of *electromagnetic energy* as also of the *wave fronts*.

Note in this context a special feature of an electromagnetic wave as compared to an acoustic wave - while the latter is essentially a scalar wave, the former is a *vector* wave involving the vectors **E** and **B**, and the variations of the two have to be related to each other according to Maxwell's equations. For a plane wave this relation is expressed, first, by the directions of the two being related to the direction of propagation of the wave as indicated above, and secondly, by the amplitudes E_0 and B_0 of the two being related as in eq. (14-5d). Additionally, the electric and magnetic field intensities, given by equations (14-5b) and (14-5c) are seen to be *in the same phase* (see sec. 14.4.3).

The phases, may however, differ in the case of wave propagation through certain media like, for instance, through a conductor.

14.4.3 The phase

This brings us to the concept of the *phase* of the wave, which is essentially the same as that introduced in chapter 9. The argument of the cosine function in eq. (14-5b) or (14-5c) is referred to as the phase (Φ ; the electric and magnetic field intensities could also be expressed in terms of the sine function whose argument could also be defined as the phase) which thus depends on x and t as

$$\Phi(x, t) = kx - \omega t. \quad (14-7a)$$

The expressions for the electric and magnetic field intensities can then be alternatively written in the form

$$E_y = E_0 \cos \Phi(x, t), \quad B_z = B_0 \cos \Phi(x, t), \quad (E_x = E_z = 0, \quad B_x = B_y = 0). \quad (14-7b)$$

Noting that both the wave functions are described in terms of the *same* value of Φ , one understands why the electric and magnetic field intensities of the wave under consideration are said to be in the same phase.

As explained in chapter 9, the phase is essentially an angular variable since values of Φ differing by integral multiples of 2π correspond to the same values of \mathbf{E} and \mathbf{B} , and one can consider only values lying in the range, say, $0 \leq \Phi < 2\pi$. It is the phase that gives us complete information as to where the instantaneous value of the wave function lies within its range of variation ($-E_0 \leq E_y \leq E_0$, $-B_0 \leq B_z \leq B_0$) and what its instantaneous rate of change is.

The wave functions (E_y and B_z) for the wave described by equations (14-5b)- (14-5d) are independent of the co-ordinates y and z depending, instead, only on x and t , its direction of propagation being along the x-axis. This is reflected in the phase Φ being a function of x and t alone. In general, however, the phase and the values of the wave functions of a monochromatic plane progressive wave depends on all the four variables x , y , z , and t , and I have chosen a particular form for the sake of simplicity.

Problem 14-2

At an instant when the electric field intensity for a plane monochromatic wave propagating along the x-axis of a right handed co-ordinate system has the maximum magnitude given by $E_0 = 0.2\text{V}\cdot\text{m}^{-1}$ at a specified point, it is directed along the z-axis of the co-ordinate system. At a point $x = 0$ and at time $t = 0$, the field intensity is $0.1\text{V}\cdot\text{m}^{-1}$, directed along the negative direction of the z-axis. If the wave, of angular frequency $6.0 \times 10^{13}\text{s}^{-1}$, is set up in a medium with $\epsilon_r = 2.5$, $\mu_r = 1.0$, find the expression for the magnetic field intensity a function of x and t .

Answer to Problem 14-2

HINT: Since the instantaneous direction of the electric field vector \mathbf{E} , that of the magnetic field vector \mathbf{B} , and the direction of propagation form a right handed triad, the required expression is of the form $\mathbf{B} = -B_0 \cos(kx - \omega t + \delta)\hat{e}_y$, while the electric field intensity is given by $\mathbf{E} = \hat{e}_z v B_0 \cos(kx - \omega t + \delta)$ (reason out why; recall that the electric field and the magnetic field are always in the same

phase); here δ is the constant part of the phase to be determined, and v is the phase velocity, for which the relations (14-5d) are satisfied. Making use of formula (14-6), we find $v = \frac{c}{\sqrt{\epsilon_r \mu_r}} = \frac{3 \times 10^8}{\sqrt{2.5}}$ m·s⁻¹, i.e., $v = 1.9 \times 10^8$ m·s⁻¹. Making use of the given maximum value of the electric field intensity, one obtains, $vB_0 = 0.2$ V·m⁻¹, i.e., $B_0 = 1.05 \times 10^{-9}$ T. The value of k is obtained as $k = \frac{\omega}{v} = \frac{6 \times 10^{13}}{1.9 \times 10^8}$ m⁻¹, i.e., 3.16×10^5 m⁻¹. Finally, making use of the value of the electric field intensity at $x = 0$, $t = 0$, one obtains $-0.1 = 0.2 \cos \delta$, i.e., $\delta = \frac{2\pi}{3}$. Thus, the required expression for the magnetic field intensity is $\mathbf{B} = -1.05 \times 10^{-9} \cos(3.16 \times 10^5 x - 6.0 \times 10^{13} t + \frac{2\pi}{3}) \hat{e}_y$ T.

14.4.4 The wave front and its propagation

Let us now consider all those points in space at any given instant of time (say, $t = 0$) at which the phase Φ equals some chosen value, say, Φ_0 . One finds that all such points lie on a *plane* $x = \frac{\Phi_0}{k}$, i.e., a plane parallel to the y- and z-axes, and perpendicular to the direction of propagation of the wave. Such a plane is referred to as a *wave front* at the given time instant, corresponding to the chosen value (Φ_0) of the phase.

At any other time instant, say, at $t = \tau$, the set of points in space corresponding to the same value (Φ_0) of the phase corresponds to another plane $x = \frac{\Phi_0}{k} + v\tau$ (check this out). One can describe this by saying that, in the time interval τ , the wave front corresponding to the phase Φ_0 travels along the x-axis with a speed v . What was earlier described as a motion of the wave profile is now seen to be equivalent to a propagation of the wave front where, in reality, the two are identical from a conceptual point of view, and the phase velocity (v) is now seen to be the velocity of propagation of the wave front, the direction of propagation being along the x-axis for the particular wave solution under consideration.

Fig. 14-1 shows a wave front, corresponding to some chosen value of the phase (say, $\Phi = \Phi_0$) at a few successive instants of time where it is seen that the wave front gets shifted from the position A at $t = 0$ to the position B at a later time ($t = \tau$) and then again to C (at $t = 2\tau$), the distance through which it gets translated being, in each case, $v\tau$. The directions of the electric and magnetic intensity vectors at any time instant

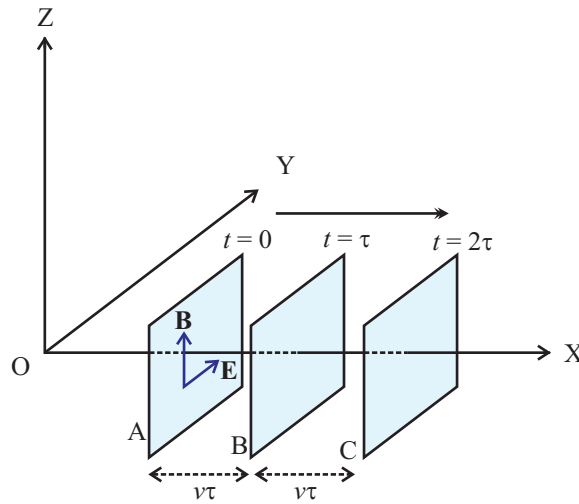


Figure 14-1: Wave fronts (schematic) of a plane progressive wave at three different instants of time ($t = 0$, $t = \tau$, and $t = 2\tau$), where τ is any chosen time interval; successive positions of the wave front corresponding to a chosen value of the phase Φ ($=\Phi_0$) are shown; the wave front is seen to propagate along the x-axis, being displaced through a distance $v\tau$ in time τ ; the double-headed arrow gives the direction of wave normal as also the direction of propagation of the wave front; directions of the electric and magnetic field vectors on the wave front are shown; other directions conforming to the right hand triad rule are also possible.

(say, at $t = 0$) are also shown, other possible directions being obtained by rigid rotations about the direction of propagation. While the two intensity vectors keep on oscillating on the wave front, one along the y-axis and the other along the z-axis in the present instance, the wave front itself propagates along the x-axis. At any given time instant, the normal to the wave front, referred to as the *wave normal*, points along the direction of propagation.

14.4.5 The electromagnetic spectrum

It may be appropriate here to explain the concept of the *electromagnetic spectrum*. Referring to plane monochromatic waves in free space, for instance, one may imagine electromagnetic waves of various possible frequencies (ν), ranging from zero to infinity (in the limiting sense), while the corresponding wavelengths (λ) in vacuum are spread over the range infinity to zero (the product $\nu\lambda$ remains constant at the value c).

Electromagnetic waves covering this entire range of frequencies or wavelengths are said to constitute the *electromagnetic spectrum*. Within this range, one may identify certain

sub-ranges such that the waves belonging to a sub-range are characterized by similar features, including the way the waves are generated, their method of detection, and the mode of their interaction with various materials. However, it may be remarked that, while the waves belonging to any one sub-range have similar characteristics, and differ somewhat from the characteristics of the waves belonging to other sub-ranges, *all* the waves making up the entire electromagnetic spectrum have a number of fundamental features in common, the most notable of which is that these are all described in terms of *Maxwell's equations* (see sec. 14.2.1.5).

As an example of these sub-ranges within the electromagnetic spectrum, I may mention *light waves* (i.e., optical waves), which constitute the *visible* part of the spectrum. This visible part covers the range of wavelengths from 350 nm to 750 nm, though the two limiting values are not sharply defined. Electromagnetic waves belonging to this visible part of the spectrum (*optical disturbances*, as they are sometimes referred to; terms like 'optical signals', and 'optical fields' are also used, sometimes with slightly different shades of meaning) are described and studied in the science of optics, especially *wave optics*.

On the lower side of the wavelength scale, as one goes to successively lower ranges of wavelength, one encounters *ultraviolet waves*, *X-rays*, and *gamma rays*. On the higher side of the wavelength scale, on the other hand, there are the *infra-red* and *heat waves*, the *microwaves*, and the *radio-waves*, belonging to sub-ranges of successively larger wavelengths.

14.4.6 Energy flux and intensity

14.4.6.1 Energy density and energy flux

It requires an expenditure of energy to set up electric and magnetic fields in any given region in space. Considering, for instance, the electric field between the plates of a parallel plate capacitor connected to the terminals of an electrical cell, one can work out the energy drawn from the cell in charging the capacitor, and interpret this as the energy stored in the electric field. Similarly, considering the magnetic field in the interior of a

solenoid drawing current from an electrical cell, one can calculate the energy required to set up the current in the solenoid and interpret it as the energy stored in the magnetic field.

More generally, we can consider the energy required to create an electric field, as with the help of a number of conductors (refer to sec 11.11.10) and, at the same time, the energy required to set up currents in a set of current-carrying wires (sec. 13.4), and interpret the sum of the two as the energy of the electromagnetic field at any given instant of time.

What one learns from these exercises is that the energy per unit volume associated with an electric field of strength \mathbf{E} at a given point in space is $u_E = \frac{1}{2}\epsilon E^2$ and that associated with a magnetic field of strength \mathbf{B} at the point is $u_M = \frac{1}{2}\frac{B^2}{\mu}$ where, for the sake of simplicity, I have assumed the medium under consideration to be an *isotropic* one, characterized by a permittivity ϵ and a permeability μ . This implies that the *energy density* at any given point in space associated with an *electromagnetic* field of strengths \mathbf{E} and \mathbf{B} at that point is

$$u = u_E + u_M = \frac{1}{2}\left(\epsilon E^2 + \frac{B^2}{\mu}\right). \quad (14-8)$$

Considering, then, any given region (say, R) in space, the total electromagnetic field energy in that region can be worked out by imagining it to be divided into a large number of tiny volume elements and summing up the energies in all those volume elements, the energy in an element of volume δv being $u\delta v$, where u stands for the energy density in the interior of the element. In the limit of the volumes of all these elements going to zero, the expression for electromagnetic field energy reduces to the integral

$$U = \int u dv, \quad (14-9)$$

where the integration extends over the volume of the given region R .

As the electric and magnetic field intensities at the various points in space vary with time, the energy of the electromagnetic field within the region R may also keep changing. The principle of conservation of energy implies that the change in field energy within R in any given time interval has to be compensated by energy changes elsewhere. Such energy changes occur on two counts - first, by the *flow* of electromagnetic energy through the boundary surface of R (see fig. 14-2) whereby the electromagnetic field energy of space outside the region R undergoes a change, and secondly, by the field doing *work* on bodies carrying charges and currents located within the region R.

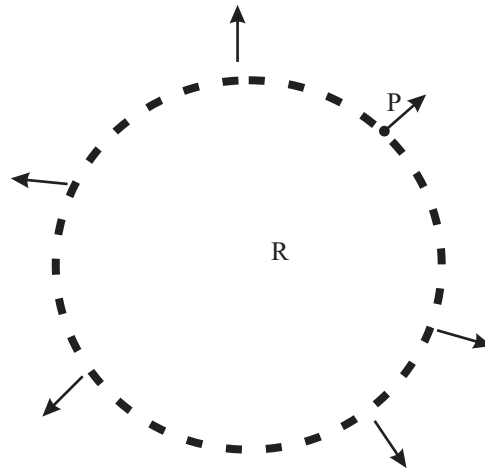


Figure 14-2: Showing a region R and the flux of energy through its boundary surface.

On carefully carrying out the energy accounting, one arrives at an expression for the *energy flux* per unit area at any given point in space. Imagining a small element of area δs around the point, oriented in a direction perpendicular to the direction of energy flow through it, if δW denotes rate of flow of energy through this element of area, then the energy flux per unit area at the point works out to

$$\mathbf{S} \equiv \frac{\delta W}{\delta s} \hat{n} = \frac{1}{\mu} \mathbf{E} \times \mathbf{B}, \quad (14-10)$$

where \hat{n} stands for a unit vector along the direction of energy flow.

This is referred to as the *Poynting vector* at the point under consideration.

Thus, the space- and time variations of the electric and magnetic intensities in an electromagnetic field result in a variation of the field energy within any given region of space, and an associated flow of energy, where the rate of flow of energy per unit area at any given point is given by the Poynting vector at that point. This flow of energy enables the field to exchange energy with bodies carrying charge and current. This means that the electromagnetic field can be looked upon as a dynamical system, similar to a system of particles or bodies in mechanics. In addition to the electromagnetic field possessing and transporting energy, it also possesses *momentum* and *angular momentum*, which it can transport from one region of space to another. What all this means is that electromagnetic theory and mechanics are similar disciplines in physics, dealing with the behaviour of *dynamical systems*.

14.4.6.2 Intensity

The energy density and the Poynting vector at any given point in space are, in general, found to vary too rapidly with time to be of any practical relevance. For one thing, if one attempts to measure the energy within a small volume in an electromagnetic field, then the finite time of response of the measuring apparatus will stand in the way of measuring the energy at any given time instant or its variation at successive instants of time. Instead, what is within the realm of practical feasibility, is a measurement of the *average* energy within the region, where the averaging is done over a finite interval of time. More generally, one is led to consider the *time-averaged values* of quantities like the energy density and the Poynting vector as being relevant from a practical point of view.

For most practical purposes, the time interval over which the averaging is to be performed so as to yield meaningful quantities of interest, is so large compared to typical time intervals characterizing the variations of the electromagnetic field vectors, that one can effectively take this to be *infinitely* large. Considering, then, a function, say f , of the

field variables, its time averaged value at any given point can be defined as

$$\langle f \rangle = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \int_0^\tau f dt, \quad (14-11)$$

where the symbol $\langle \cdot \rangle$ is used to denote the time averaging.

If the field vectors vary harmonically with time, with an angular frequency ω and a time period $T = \frac{2\pi}{\omega}$, then the above expression is equivalent to

$$\langle f \rangle = \frac{1}{T} \int_0^T f dt. \quad (14-12)$$

With this interpretation of time averaging, quantities like the energy density and the Poynting vector will, from now on, be meant to refer to the time-averaged values like $\langle u \rangle$ and $\langle \mathbf{S} \rangle$ respectively.

The magnitude of the (time-averaged) Poynting vector ($|\langle \mathbf{S} \rangle|$) at any given point is termed the *intensity* at that point.

For the plane monochromatic wave represented by equations (14-5b) and (14-5c), the intensity is the same at all points in space and is given by the expression

$$I = \sqrt{\frac{\epsilon_r \epsilon_0}{\mu_r \mu_0}} \frac{E_0^2}{2}. \quad (14-13a)$$

For a plane monochromatic wave propagating through vacuum, this reduces to

$$I = \sqrt{\frac{\epsilon_0}{\mu_0}} \frac{E_0^2}{2} = N E_0^2 \text{ (say)}, \quad (14-13b)$$

where N is a constant of proportionality that one may, in a given context, set equal to unity while *comparing* intensities at different points in space.

The expression (14-13b) for the intensity holds for a plane monochromatic wave propagating in any given direction (specified by the unit vector, say, \hat{n} ; see sec. 14.4.9) in

space: *the intensity is proportional to the squared modulus of the amplitude of the electric field vector*, the constant of proportionality being $\frac{1}{2}\sqrt{\frac{\epsilon}{\mu}}$, where ϵ and μ stand for the permittivity and permeability of the medium through which the wave propagates.

Making use of the complex representation of a plane electromagnetic wave (see sec. 14.5 below) and denoting the complex amplitude of the electric intensity vector at any given point by $\tilde{\mathbf{E}}$, the relative intensity at that point can be expressed in the form $N\langle|\tilde{\mathbf{E}}|^2\rangle$, where the symbol $\langle\cdot\rangle$ stands for time averaging.

14.4.6.3 Velocity of energy transport

At the same time, one can work out the time-averaged energy density of the electromagnetic field at any given point which, for a plane monochromatic wave, works out to

$$\langle u \rangle = \frac{1}{2}\epsilon E_0^2. \quad (14-14)$$

Once again, this expression holds for a plane monochromatic wave propagating in any given direction in space (see problem 14-3).

Comparing expressions (14-13a) and (14-14), one obtains

$$I = v\langle u \rangle, \quad (14-15)$$

where $v = \frac{1}{\sqrt{\epsilon\mu}}$ (see eq. (14-6)) is the phase velocity of the plane wave in the medium under consideration. The phase velocity is related to the refractive index n of the medium by the third third in (14-5d).

This result can be interpreted as implying that *the velocity of energy transport by a plane monochromatic electromagnetic wave is the phase velocity v* . One can see this as follows.

For any given point P and a given time interval δt , imagine a cylindrical volume element

around P (see fig. 14-3) with its axis along the direction of the time-averaged Poynting vector, where the length of the cylinder is $v\delta t$ and the area of cross-section is, say, δA . The electromagnetic field energy within this cylinder is $\langle u \rangle v \delta t \delta A$. Assuming that v does indeed represent the velocity of energy flow, this amount of energy will flow out of the cylinder in the given time interval δt . The intensity at P which, by definition, is the time-averaged rate of flow of energy per unit area, will then be $\frac{\langle u \rangle v \delta t \delta A}{\delta t \delta A} = v \langle u \rangle$, which is precisely eq. (14-15). This tells us that our assumption of v being the velocity of energy transport by the plane wave is indeed a valid one.

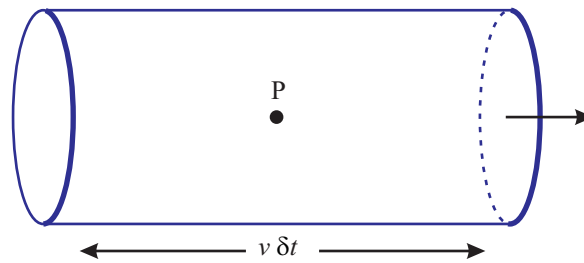


Figure 14-3: A cylindrical volume element around a point P, with the length of the cylinder along the direction of flow of energy at P.

1. A perfectly monochromatic plane wave is an idealization. From a practical point of view, it is more appropriate to consider a quasi-monochromatic wave which is obtained by a superposition of monochromatic waves with frequencies lying within some small range. This results in what can be termed a *wave packet* that propagates in space with a velocity referred to as the *group velocity* (v_g) of the medium. In general, the group velocity differs from the phase velocity of a monochromatic wave due to the dependence of the latter on the frequency - a phenomena known as *dispersion* (see sec. 14.7).

For such a quasi-monochromatic wave, the velocity of transport of electromagnetic field energy is no longer the phase velocity v , but is actually given by the *group velocity* v_g .

2. The results stated above hold for plane monochromatic waves propagating through an isotropic medium which is, moreover, a *non-absorbing* one. In the case of a wave propagating through a conducting medium, for instance, some of the ex-

pressions get modified.

Problem 14-3

Establish formula (14-14) for a monochromatic plane wave propagating along any specified direction in space.

Answer to Problem 14-3

HINT: Referring to formula (14-8), the instantaneous energy density for a plane wave propagating along any specified unit vector \hat{n} in a medium with permittivity and permeability ϵ and μ respectively is obtained by putting $\mathbf{E} = \hat{l}E_0 \cos(kx - \omega t + \delta)$, $\mathbf{B} = \hat{m}B_0 \cos(kx - \omega t + \delta)$, where \hat{l} and \hat{m} are unit vectors perpendicular to each other, and $\hat{l} \times \hat{m} = \hat{n}$, $B_0 = \frac{E_0}{v}$ (v = phase velocity of the wave = $\frac{1}{\sqrt{\epsilon\mu}}$), ω is the angular frequency ($= \frac{2\pi}{T}$, T = time period of oscillation of the electric and magnetic field vectors), $k = \frac{\omega}{v}$, and δ denotes the constant part of the phase. The time averaged energy density is $\langle u \rangle = \frac{1}{T} \int_0^T u dt$. For any specified value of x , one has the result $\frac{1}{T} \int_0^T \cos^2(kx - \omega t + \delta) = \frac{1}{2}$. Hence, $\langle u \rangle = \frac{1}{4}(\epsilon E_0^2 + \frac{1}{\mu} B_0^2)$. Now, $\frac{1}{\mu} B_0^2 = \frac{1}{v^2 \mu} E_0^2 = \epsilon E_0^2$. This gives $\langle u \rangle = \frac{1}{2} \epsilon E_0^2$, as required.

Problem 14-4

The maximum magnitude of the electric field intensity at a point 15 m away from a point source of light in free space is $E_0 = 1.5 \text{ V}\cdot\text{m}^{-1}$. What is the maximum value of the magnetic field intensity and the intensity of light at that point? calculate the power radiated by the source.

Answer to Problem 14-4

HINT: At large distances from a point source (in the present instance, the distance 15 m can indeed be considered large compared to the wavelength of light) the electric and magnetic field vectors behave locally like those in a plane wave. Hence the maximum magnetic field intensity $B_0 = \frac{E_0}{c} = 0.5 \times 10^{-8} \text{ T}$. The intensity is then (see eq. (14-13a)) $I = \sqrt{\frac{\epsilon_0}{\mu_0}} \frac{E_0^2}{2} = 2.98 \times 10^{-3} \text{ W}\cdot\text{m}^{-2}$. The power P radiated by the source is transported uniformly through the surface of the sphere with radius $r = 15 \text{ m}$, which gives $P = 4\pi r^2 I = 8.43 \text{ W}$.

14.4.7 Radiation pressure

With reference to a plane electromagnetic wave propagating (say, through free space) in any given direction, a fact of basic importance is that, there takes place not only a transport of energy by means of the wave, but a transport of *momentum* as well. In other words, the electromagnetic field can be looked upon as a dynamical system like, say, a particle or a system of particles with which one can associate an energy as well as a momentum. Thus, considering any point P in a region of space through which the wave propagates, and imagining a small area δs around that point oriented in such a manner that the direction of propagation of the wave is perpendicular to it, there occurs a flow of momentum (where only the component of momentum along the direction of propagation of the wave is involved), say, δP per unit time through that area. The time average of the *rate of flow of momentum per unit area*, i.e., $\frac{\delta P}{\delta s}$, then represents the *pressure* due to the plane wave at P.

In the simple situation corresponding to the propagation of a plane electromagnetic wave in vacuum, the pressure is given by the formula

$$p = \frac{I}{c}, \quad (14-16a)$$

where I stands for the intensity at the point under consideration, and c is the velocity of the electromagnetic wave in vacuum. Alternative expressions for the pressure are then

$$p = u = \frac{1}{2} \epsilon_0 E_0^2, \quad (14-16b)$$

where u stands for the time averaged energy density and E_0 for the amplitude of the electric field intensity at the point under consideration.

Experimentally, the momentum transport by means of the electromagnetic wave can be made evident by letting the wave hit an obstacle, when the latter gains momentum from the wave in a manner analogous to the momentum transferred in a collision between two bodies. If the wave is completely absorbed by the obstacle, the rate of gain of momentum by the obstacle is $pA = \frac{IA}{c}$, where A stands for the normal (i.e., perpendicular to the

direction of propagation of the wave) area presented by the obstacle to the wave. If, on the other hand, the wave is incident on an obstacle that acts as a perfect *reflector* of the wave, the pressure felt by the obstacle will be $p' = 2p = \frac{2I}{c}$, and the force experienced by it will be $\frac{2IA}{c}$.

In contrast to a plane wave where energy and momentum are transported in a specific direction, namely, the direction of propagation of the wave, the electromagnetic waves making up *black body radiation* (see sections 8.23.3.1, 16.9) within an enclosure propagate *isotropically* in all directions. the pressure of black body radiation at any point is therefore given by the formula

$$p = \frac{u}{3}, \quad (14-17)$$

where u once again stands for the time averaged energy density at the point under consideration.

Problem 14-5

A plane unpolarized electromagnetic wave with intensity $12 \text{ mW} \cdot \text{m}^{-2}$ is sent through a polaroid sheet (a thin layer of a polarizing material, whose molecules are all oriented in such a manner that the sheet can transmit only those electromagnetic waves which have their electric field vector oscillating along some specific direction), giving rise to a linearly polarized wave. Find the maximum value of the electric field intensity of the resulting beam, and the radiation pressure experienced by the polaroid sheet.

Answer to Problem 14-5

Since unpolarized light contains an equal mixture of two types of linearly polarized light with electric field vectors oscillating in mutually perpendicular directions (see sec. 14.4.8 below for background), the intensity associated with any one of the two components has to be half the intensity of unpolarized light. Thus, the intensity transmitted by the polaroid sheet is $6 \text{ mW} \cdot \text{m}^{-2} = \sqrt{\frac{\epsilon_0}{\mu_0}} \frac{E_0^2}{2}$ (refer to eq. (14-13a), where we assume that the wave propagates in free space). This gives $E_0 = 2.13 \text{ V} \cdot \text{m}^{-1}$.

Of the two linearly polarized components making up the unpolarized wave, one is transmitted by the polaroid while the other is absorbed by it. Hence the pressure experienced by the polaroid is the rate of momentum transported by the latter component per unit area, i.e., $p = \frac{I}{c}$, where I stands for the intensity of either of the two linearly polarized components ($6 \text{ mW} \cdot \text{m}^{-2}$). This gives, $p = 2.0 \times 10^{-11} \text{ Pa}$.

14.4.8 The state of polarization of an electromagnetic wave

The wave described by equations (14-5b)- (14-5d) is referred to as a *linearly polarized* (or, alternatively, *plane polarized*) wave. This corresponds to the fact that the electric field intensity vector at any point in space oscillates along a fixed direction, namely, the y-axis in this instance (a similar statement being valid for the magnetic vector as well, which oscillates along the z-axis).

More generally, a linearly polarized monochromatic plane wave propagating along the x-axis can have its electric vector oscillating along any other fixed direction in the y-z plane, in which case its magnetic vector will oscillate along a perpendicular direction in the same plane (recall that for a plane progressive wave the electric vector, the magnetic vector, and the direction of propagation have to form a right handed orthogonal triad - a requirement imposed by Maxwell's equations). For instance, one can think of a linearly polarized plane monochromatic wave propagating in the x-direction, where the directions of oscillation of the electric and magnetic intensities in the y-z plane are as shown in fig. 14-4.

Such a linearly polarized wave can be looked upon as a superposition of two constituent waves, each linearly polarized, the phase difference between the two waves being zero. More precisely, consider the wave described by equations (14-5b)- (14-5d), and call it the *first wave*. Consider, in addition, a *second* wave, also a solution of Maxwell's equations

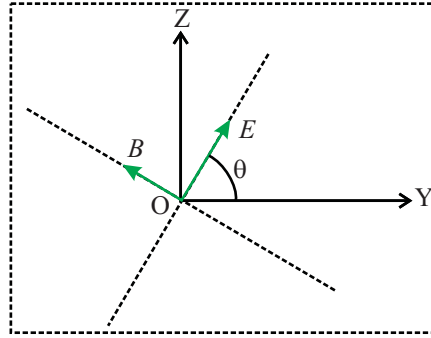


Figure 14-4: Depicting the directions of oscillation (dotted lines) of the electric and magnetic field vectors of a linearly polarized plane progressive wave propagating along the x-axis (perpendicular to the plane of the figure, coming out of the plane), where the direction of the electric field intensity is inclined at an angle θ with the y-axis; correspondingly, the direction of the magnetic vector is inclined at the same angle with the z-axis, the two vectors being shown at an arbitrarily chosen instant of time; the wave is obtained by a linear superposition (equations (14-19a), (14-19b)) of two linearly polarized waves, one with the electric vector oscillating along the y-axis and the other with the electric vector oscillating along the z-axis, the phases of the two waves being the same.

for the medium under consideration, described by the equations

$$E'_z = E_0 \cos(kx - \omega t) \quad (E'_x = E'_y = 0), \quad (14-18a)$$

$$B'_y = -B_0 \cos(kx - \omega t) \quad (B'_x = B'_z = 0), \quad (14-18b)$$

along with equations (14-5d).

Denoting the electric and magnetic vectors at any point \mathbf{r} and at time t for the first and second waves by $\mathbf{E}(\mathbf{r}, t)$, $\mathbf{B}(\mathbf{r}, t)$ and $\mathbf{E}'(\mathbf{r}, t)$, $\mathbf{B}'(\mathbf{r}, t)$ respectively, consider the wave for which the electric and magnetic vectors are described by the following superposition of these two waves, with amplitudes taken in the ratio $\cos\theta : \sin\theta$ (where the angle θ may be chosen arbitrarily)

$$\bar{\mathbf{E}}(\mathbf{r}, t) = \cos\theta \mathbf{E}(\mathbf{r}, t) + \sin\theta \mathbf{E}'(\mathbf{r}, t), \quad (14-19a)$$

$$\bar{\mathbf{B}}(\mathbf{r}, t) = \cos\theta \mathbf{B}(\mathbf{r}, t) + \sin\theta \mathbf{B}'(\mathbf{r}, t). \quad (14-19b)$$

Note that the phases of the two waves under consideration are the same (we assume, for the sake of concreteness, $0 \leq \theta \leq \frac{\pi}{2}$; negative values of θ are also possible, see below). For instance, E_y and E'_z are characterized by the same phase ($\Phi(x, t) = kx - \omega t$) and so are B_z and $-B'_y$ (where the minus sign indicates that the reference direction for \mathbf{B}' is to be taken along the unit vector $-\hat{e}_y$, as required by the right hand rule for the electric field intensity, the magnetic field intensity, and the direction of propagation). This resultant wave then corresponds to a linear polarization as shown in fig. 14-4. A phase difference of π between the two superposed waves also results in a linearly polarized wave, but now with the direction of oscillation of the electric field vector making an angle $-\theta$ with the y-axis.

Since the amplitude of oscillation of the electric field intensity for the superposed wave is the same as either of the two constituent waves, i.e., E_0 (check this out), the intensity of the superposed wave is once again given by the expression (14-13a).

More generally, one may consider superpositions of two linearly polarized waves (with mutually perpendicular directions of oscillation of the electric field vector) with a constant phase difference *other* than zero or π . Such a superposition results, in general, to what is termed an *elliptically* polarized wave.

From a practical point of view, two waves with a constant phase difference between them, may arise from two sources *correlated* to each other in a definite manner. On the other hand, if the two sources are *uncorrelated* to each other then it is more likely that the radiation from the source is made up of a large number of such superpositions where the phase difference cannot be assigned a definite value, but is described by a *random* variable. Such a mixture of superposed waves would then correspond to an *unpolarized* wave. Elliptically polarized waves (special instances of which correspond to linearly polarized and *circularly* polarized waves) and unpolarized waves constitute two extreme cases of *partially* polarized waves that can be looked upon as mixtures of polarized and unpolarized ones.

In summary, the vector nature of an electromagnetic wave tells us that a complete description of a plane progressive wave has to involve the specification of its *state of polarization* since the wave may be linearly polarized or, more generally, elliptically polarized. A still more general type of wave disturbance is an *unpolarized* wave which is obtained from a superposition of two linearly polarized plane progressive waves with a *randomly* fluctuating phase difference between the two. Such a wave disturbance is termed *incoherent*, in contrast to a linearly or elliptically polarized monochromatic plane wave which corresponds to a *coherent* wave disturbance. We will be having a brief look at incoherent wave disturbances in section 14.10.

A particular instance of elliptic polarization is a *circularly polarized wave* which is made up of two linearly polarized waves with mutually perpendicular directions of polarization as considered above (with amplitudes taken in the ratio 1 : 1, i.e., with $\theta = \frac{\pi}{4}$ in equations (14-19a), (14-19b)), superposed with a phase difference of $\frac{\pi}{2}$ or $\frac{3\pi}{2}$ between them. Circularly polarized plane progressive waves are of great importance in present day *optical information processing*.

I have talked of polarization in the context of plane progressive electromagnetic waves in this section. However, the concept of polarization extends to electromagnetic waves of certain other descriptions as well where the directions of oscillations of the electric and magnetic field vectors bear a definite and characteristic relationship with the direction of propagation of the wave. Instances where a wave can be characterized in such a manner are what are known as the transverse magnetic (TM) and transverse electric (TE) spherical waves (see section 14.8 for an introduction to spherical and cylindrical waves) in regions of space far from their sources. Similar characterizations are also possible for a class of cylindrical waves as well. However, I will not enter here into a detailed description and analysis of these waves.

14.4.9 Wave propagating in an arbitrarily chosen direction

While equations 14-5b and 14-5c describe a plane progressive wave propagating along the x-axis (of a chosen co-ordinate system), the expressions can be generalized to de-

scribe a wave propagating in any arbitrarily chosen direction. Thus, let \hat{n} be a unit vector in the direction chosen and \hat{m} be a unit vector in any chosen direction in the plane perpendicular to \hat{n} . Then, a linearly polarized plane progressive wave propagating in the direction of \hat{n} and polarized with the electric field vector oscillating along a line parallel to \hat{m} is given by the equations

$$\mathbf{E} = \hat{m}E_0 \cos(k(\hat{n} \cdot \mathbf{r}) - \omega t), \quad (14-20a)$$

$$\mathbf{B} = \frac{1}{c} \hat{n} \times \mathbf{E}, \quad (14-20b)$$

where the arguments \mathbf{r} and t have been suppressed in \mathbf{E} and \mathbf{B} . Once again, the electric field intensity, the magnetic field intensity, and the direction of propagation form a right-handed orthogonal triad. The vector $\mathbf{k} = k\hat{n}$ is referred to as the *wave vector* of the plane wave.

Considering any plane perpendicular to \hat{n} , the equation of the plane can be written in the form $\hat{n} \cdot \mathbf{r} = \text{constant}$. According to the definition of a wave front, this means that any such plane is a wave front for the wave given by the above expressions, and \hat{n} therefore gives the direction of a *wave normal* for the plane wave. It is, moreover, seen from the above expressions that the electric and magnetic field vectors at any given point lie in the wave front passing through that point, and are orthogonal to the wave normal. All these features are, in reality, re-statements of what we found for a plane wave given by equations (14-5b) and (14-5c).

Monochromatic plane progressive waves can be generated by special arrangements involving monochromatic sources and *collimating* devices like lenses or parallel mirrors. In optics, a *laser source* produces a very good approximation to a monochromatic plane wave.

The plane progressive waves described in this section and earlier in section 14.4 are very special in that they belong to the class of *coherent* waves (see section 14.10), while wave disturbances are, more generally, *incoherent* or *partially coherent* in practice. More-

over, a monochromatic wave is also an idealization in that waves are, more generally, *polychromatic*, i.e., made up of a number of monochromatic components with different frequencies. In spite of these idealized features, plane progressive monochromatic waves are of great theoretical and practical importance in electromagnetic theory and optics.

For instance, the plane progressive waves are *basic* in the sense that other coherent electromagnetic disturbances can be expressed as superpositions of these simple wave solutions of the Maxwell equations. Incoherent electromagnetic disturbances can also be looked upon as being made up of ‘superpositions’ (more appropriately termed ‘mixtures’) of plane progressive waves whose amplitudes, frequencies, and phases are randomly distributed.

14.4.10 Wave normals and rays

From the definition of the Poynting vector (eq. (14-10)), one finds that the direction of energy transport at any given point for the plane wave described by equations (14-20a) and (14-20b) is along the wave normal passing through that point. In other words, the wave normals for a plane wave form a family of parallel straight lines along which the transport of electromagnetic field energy takes place. Such paths of propagation of the field energy are referred to as *ray paths*. Thus, in summary, the ray paths for a plane wave are simply the wave normals.

Problem 14-6

The magnetic field intensity at a point \mathbf{r} and at time t for a monochromatic plane wave is given by $B_0 \cos \omega(\frac{1}{c}\hat{n} \cdot \mathbf{r} - t)\hat{m}$, where $\hat{n} = \frac{1}{\sqrt{3}}(\hat{i} + \hat{j} + \hat{k})$ (\hat{i} , \hat{j} , \hat{k} are the unit vectors along the three axes of a right handed co-ordinate system), $B_0 = 2.0 \times 10^{-9}$ T, $\omega = 4 \times 10^{13}$ s $^{-1}$, and $\hat{m} = \frac{1}{\sqrt{6}}(\hat{i} - 2\hat{j} + \hat{k})$. Find an expression for the electric field vector as a function of \mathbf{r} , t and obtain the time averaged energy density at any point. Work out the energy flowing through a unit area, held perpendicular to the x-axis, around any given point in a complete period of oscillation of the electric and magnetic field intensities.

Answer to Problem 14-6

HINT: Since the electric field intensity, magnetic field intensity, and the direction of propagation form a right handed triad, the unit vector along the direction of \mathbf{E} , when that along \mathbf{B} is along \hat{m} , is $\hat{m} \times \hat{n} = \frac{1}{\sqrt{2}}(-\hat{i} + \hat{k})$. Further, the maximum magnitude of the electric field intensity is given by $E_0 = cB_0 = 0.6 \text{ V}\cdot\text{m}^{-1}$. In other words, the required expression for the electric field vector is $\mathbf{E}(\mathbf{r}, t) = \frac{0.6}{\sqrt{2}} \cos\left(4 \times 10^{13} \left(\frac{1}{3\sqrt{3} \times 10^8}(x + y + z) - t\right)\right)(-\hat{i} + \hat{k}) \text{ V}$. The time averaged energy density is $u = \frac{1}{2} \epsilon_0 E_0^2$ (refer to formula (14-14)), i.e., $u = \frac{1}{2} \times 8.85 \times 10^{-12} \times (0.6)^2 \text{ J}\cdot\text{m}^{-3}$, i.e., $1.6 \times 10^{-12} \text{ J}\cdot\text{m}^{-3}$ (approx). The intensity, which is the time-averaged rate of flow of energy per unit area held perpendicular to the direction of wave normal (which is along \hat{n}) is $I = cu$. Since the x-axis is inclined at an angle of $\theta = \arccos \frac{1}{\sqrt{3}}$ to the direction of wave normal, and since the time average refers to a period of one complete oscillation ($T = \frac{2\pi}{\omega}$), the energy flowing through a unit area, held perpendicular to the x-axis in one complete period is $U = I \cos \theta T = 3 \times 10^8 \times 1.6 \times 10^{-12} \times \frac{1}{\sqrt{3}} \times \frac{2\pi}{4 \times 10^{13}} \text{ J}\cdot\text{m}^{-2}$, i.e., $4.35 \times 10^{-17} \text{ J}\cdot\text{m}^{-2}$.

14.5 The complex representation of wave functions

While talking of simple harmonic motion in chapter 5, we took a broad view of things and looked at simple harmonic oscillations of physical quantities of various different kinds. For instance, the oscillations of currents and voltages in AC circuits constitute a particular example of simple harmonic oscillations where, in general, damping and forcing terms may also be present in the equation of motion. Broadly speaking, natural and forced simple harmonic oscillations involve a sinusoidal variation of a physical quantity with time, as in the steady state oscillations of a particle subjected to a periodic forcing (see section 4.6). As indicated in section 13.5.1.3, it is often convenient to represent such sinusoidal variations in terms of *complex quantities* where it is understood that physical quantities of relevance correspond to the *real* (or, alternatively, the imaginary) parts of these complex representations.

This approach of making use of complex representations may be extended to the analysis and description of *wave motions* as well, and is widely employed in electromagnetic

theory and optics, as also in acoustics. As we have seen, a wave is described by the spatial and temporal variations of a *wave function*, where these variations occur in accordance with a *wave equation*, simple solutions of the latter being a monochromatic plane wave, a spherical wave, or a cylindrical wave. These simple solutions are relevant in the context of more complex solutions of the wave equation in that the latter may be built up from a *superposition* of some of these simpler solutions.

A monochromatic plane wave propagating along the x-axis may be represented in the form

$$\psi = A \cos(kx - \omega t + \delta), \quad (14-21)$$

where, for instance, the wave function ψ may stand for the excess pressure in an acoustic wave (refer to equation (9-6)), or a component of the electric or magnetic field intensity in an electromagnetic wave (equations (14-5b) and (14-5c)), and where a constant phase δ has been introduced for the sake of generality.

The *complex representation* of the wave function then looks like

$$\tilde{\psi} = \tilde{A} e^{i(kx - \omega t)}, \quad (14-22)$$

where $\tilde{A} = A e^{i\delta}$ is the complex amplitude of the plane wave and $\Phi = kx - \omega t$ is the phase of the wave, now appearing through the complex phase factor $e^{i\Phi}$. The real and the complex representations of the wave are related as

$$\psi = \text{Re}(\tilde{\psi}). \quad (14-23)$$

As indicated in section 13.5.1.3, the complex representation leads to a considerable advantage in the mathematics relating to waves and oscillations.

In numerous situations of interest, the complex representation of a wave function looks

like

$$\psi(\mathbf{r}, t) = \tilde{A}(\mathbf{r})e^{i(\phi - \omega t)}, \quad (14-24)$$

where the variation with time is assumed to be a harmonic one with angular frequency ω and where the complex amplitude \tilde{A} depends only on the position \mathbf{r} . In a complex representation of this form, ϕ stands for the space dependent part of the phase since the time dependent part has been included separately. What is of considerable relevance is the fact that, in these situations, ϕ involves a factor $k = \frac{\omega}{v}$, where v stands for the velocity of propagation of the wave. For electromagnetic waves of short wavelength, including optical waves, k happens to be a *large* quantity, as a result of which, the *spatial* variation of the wave function occurs principally through ϕ , and not through \tilde{A} . It is this fact that enables one to make intelligent approximations whereby the description of a wave becomes relatively simple. An instances of such simplification relates to the description of the propagation of light with the help of *rays*.

Simple instances of electromagnetic fields conforming to a description of the form (14-24) are the ones corresponding to a plane progressive monochromatic wave, and to monochromatic spherical and cylindrical waves (see section 14.8 for an introduction to spherical and cylindrical waves), the latter two in regions of space far from their sources.

Incidentally, while making use of the complex representation, the tilde on top of the relevant symbols is sometimes suppressed so as not to make the notation look too cumbersome. Explicit mention is made of the intended meanings of the symbols whenever there is scope of confusion.

14.6 Reflection and refraction of plane waves

Fig. 14-5 depicts schematically a plane wave incident on the plane interface separating two homogeneous media (say, A and B) where a co-ordinate system is chosen with the

plane interface lying in its x-y plane, and the normal to the interface at any point on it being along the z-axis. The figure shows a wave normal intersecting the interface at O, where the wave normal can be described, for the plane wave under consideration, as a ray incident at O (see sec. 14.4.10). The wave front is then perpendicular to the ray, with the electric and magnetic field vectors oscillating in the plane of the wave front. The plane of the figure, containing the ray and the normal to the surface at O, is the x-z plane of the co-ordinate system chosen and the unit vector along the direction of the ray is, say,

$$\hat{n} = \hat{i} \cos \theta + \hat{k} \sin \theta, \quad (14-25)$$

where \hat{i} and \hat{k} denote unit vectors along the x- and z-axes, and θ is the angle made by the ray with the interface, i.e., in the present case, with the x-axis.

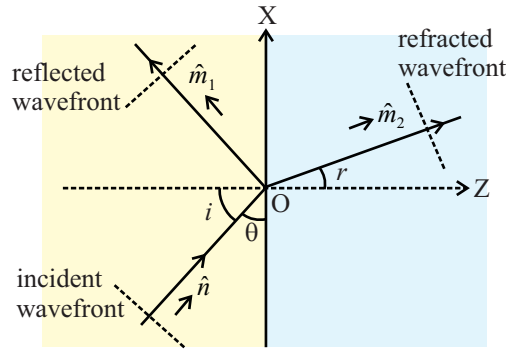


Figure 14-5: Plane wave incident on a plane interface separating two media: illustrating the laws of reflection and refraction; a wave incident on the interface with its wave normal along \hat{n} gives rise to a reflected wave and a refracted one, with wave normals along \hat{m}_1 and \hat{m}_2 respectively; the three wave normals have to be geometrically related in a certain manner (laws of reflection and refraction) so that a certain set of boundary conditions can be satisfied on the interface.

The complex representation of the wave function describing the wave (see sections 14.4.9 and 14.5) is then of the form

$$\psi = A \exp \left[i \left(\frac{2\pi}{\lambda_1} (x \cos \theta + z \sin \theta) - \omega t \right) \right], \quad (14-26)$$

where ω stands for the angular frequency of the wave (which we assume to be monochro-

matic), and λ_1 for its wavelength in the medium of incidence (i.e., medium A). For the sake of simplicity, we consider here a scalar wave function ψ which, in the present instance, can be chosen to be any of the Cartesian components of the field vectors. Finally, A , which we choose to be real and positive, stands for the amplitude of the wave.

In eq. (14-26), the angular frequency and the wavelength λ_1 are related to each other as

$$\omega\lambda_1 = \frac{2\pi c}{n_1}, \quad (14-27)$$

where c is the velocity of light in vacuum, and n_1 is the refractive index of the medium A. One arrives at eq. (14-26) from (14-20a) or (14-20b) by making use of eq (14-25).

At the interface, the wave under consideration suffers reflection and refraction, where the reflected and refracted plane waves propagate in the media A and B respectively. Let us assume that the wave normals for the reflected and refracted plane waves point along the unit vectors \hat{m}_1 and \hat{m}_2 respectively. One can then write down the expressions for the complex representations of the wave functions for these two waves in the forms

$$\psi_1 = A_1 \exp \left[i \left(\frac{2\pi}{\lambda_1} \hat{m}_1 \cdot \mathbf{r} - \omega t \right) \right], \quad (14-28a)$$

$$\psi_2 = A_2 \exp \left[i \left(\frac{2\pi}{\lambda_2} \hat{m}_2 \cdot \mathbf{r} - \omega t \right) \right], \quad (14-28b)$$

respectively, where now the amplitudes A_1 and A_2 can, in principle, be complex. Of crucial relevance here are the expressions of the space-dependent parts of the *phases* of the three waves, which we write as

$$\phi = \frac{2\pi}{\lambda_1} (x \cos \theta + z \sin \theta), \quad \phi_1 = \frac{2\pi}{\lambda_1} \hat{m}_1 \cdot \mathbf{r}, \quad \phi_2 = \frac{2\pi}{\lambda_2} \hat{m}_2 \cdot \mathbf{r}, \quad (14-29)$$

for the incident wave, the reflected wave, and the refracted, or transmitted, wave respectively. The wave disturbance in medium A in the situation under consideration can be described as a superposition of ψ and ψ_1 , while that in medium B is described by ψ_2 .

These waves in the two media have to satisfy a set of *boundary conditions* at the surface of separation, these conditions being necessary so as to make Maxwell's equations in the two media consistent with each other. These boundary conditions relate the components of the field vectors on either side of the interface, as any given point on the interface is approached from the two sides. Translated in terms of the phases, this requires

$$\phi = \phi_1 = \phi_2, \quad (14-30)$$

where now ϕ, ϕ_1, ϕ_2 are to be evaluated at any arbitrarily chosen point, say $(x_0, y_0, 0)$ on the interface (recall that the interface coincides with the x-y plane).

14.6.1 Reflection

The relation $\phi = \phi_1$ in (14-30) gives

$$x_0 \cos \theta + z_0 \sin \theta = m_{1x} x_0 + m_{1y} y_0, \quad (14-31)$$

which is to hold for all values of x_0, y_0 corresponding to points lying on the interface. One thereby obtains

$$m_{1x} = \cos \theta, m_{1y} = 0. \quad (14-32)$$

The second of the above two relations implies that the wave normal of the reflected wave is perpendicular to the y-axis, i.e., in other words, lies in the plane defined by the incident wave normal and the normal to the interface. The first relation also has a simple and familiar interpretation. Referring to fig. 14-5, one has

$$m_{1x} = \sin i', \cos \theta = \sin i, \text{ i.e., } i = i', \quad (14-33)$$

where i and i' stand for the angles of incidence and reflection respectively.

These one recognizes as the *laws of reflection* familiar in ray optics.

14.6.2 Refraction

In a similar manner, the relation $\phi = \phi_2$ in (14-30) leads to

$$m_{2y} = 0, \quad \frac{\cos\theta}{\lambda_1} = \frac{m_{2x}}{\lambda_2}. \quad (14-34)$$

The first of these relation implies that the incident wave normal, the refracted wave normal, and the normal to the refracting surface are coplanar, i.e., in other words, the incident ray, the refracted ray, and the normal to the refracting surface at the point of incidence lie in the same plane. As for the second relation in (14-34), note that, analogous to eq. (14-27), one has

$$\omega\lambda_2 = \frac{2\pi c}{n_2}, \quad (14-35)$$

where n_2 stands for the refractive index of medium B. This, then, implies

$$\frac{\sin i}{\sin r} = \frac{n_2}{n_1}. \quad (14-36)$$

In other words, the principles of electromagnetic theory lead to the *laws of refraction* as well.

There exist an infinite number of wave normals, all parallel to one another, for a plane wave. Of these, one particular member is shown in fig. 14-5, corresponding to a ray path incident at the point O. Similarly, the reflected and refracted ray paths correspond to one particular wave normal each for the reflected and refracted wave fronts, where the ray paths are determined by the requirement that they have to pass through the point O. This allows us to make statements relating to the incident, reflected, and refracted rays, instead of ones relating to just the directions of wave normals that equations (14-33) and (14-36) imply.

14.6.3 Total internal reflection

In the case of refraction from an optically denser to a rarer medium ($n_1 > n_2$) eq. (14-36) implies that the angle of refraction equals $\frac{\pi}{2}$, the largest possible value that it can have, for

$$i = i_C \equiv \sin^{-1} \mu, \quad (\mu \equiv \frac{n_2}{n_1}), \quad (14-37)$$

where μ stands for the refractive index of the medium B *relative* to the medium A. The angle i_C is referred to as the critical angle of incidence for the two media, since, for $i > i_C$, there is no refracted wave (of the form given by eq. (14-28b)) sent into the second medium (medium B; for a refracted wave to exist, $\sin r$ would have to be less than unity). One then has only the reflected wave in the first medium (medium A), i.e., the entire energy carried by the incident wave is sent back to the first medium (strictly speaking, one should speak here of the energy propagating in a direction perpendicular to the refracting surface; there occurs an energy flow parallel to the surface as well, but it involves both the two media for all angles of incidence; see sec. 10.6).

This is the phenomenon of *total internal reflection* familiar in ray optics, which we here find to follow from the principles of electromagnetic theory. However, electromagnetic theory leads to a number of additional conclusions relating to total internal reflection that do not appear in the ray theoretic description of the phenomenon.

A result of crucial relevance that follows from the principles of electromagnetic theory applied to the situation under consideration tells us that, in spite of the fact that there is no refracted ray in the second medium, *the electric and magnetic field vectors in this medium are non-zero*, i.e., a wave disturbance does enter into this medium, though it is not of such a kind as to correspond to a flow of energy through it in a direction normal to the interface. This wave disturbance in the second medium belongs to the category of an *inhomogeneous wave*, where the wave propagates in a direction parallel to the interface, but its amplitude decreases exponentially with the distance (z) from it.

Since ray optics is just an approximate description of a certain class of phenomena

(in this context, see sec. 15.9) for which electromagnetic theory gives a more complete description, these features relating to total internal reflection are not captured in a ray-theoretic analysis.

14.7 Dispersion and absorption

The plane wave solution we have considered is simple in the sense that it propagates with a constant amplitude. In reality, however, the amplitude decreases as the wave propagates through a medium, and the intensity also decreases accordingly. This is the phenomenon of *absorption*. Moreover, considering plane waves of various different frequencies, the phase velocity is found to change with the frequency. This is the phenomenon of *dispersion*. And the reason I am talking of absorption and dispersion together is that the two are, in general, related.

14.7.1 Plane waves in a dielectric medium

When an electromagnetic wave propagates through a dielectric medium, the oscillations of the electric field intensity results in tiny oscillating dipoles being produced throughout the medium. The dipole moment per unit volume around any given point is termed the *polarization* vector at that point. One can work out the relation between the polarization vector and the electric field intensity of the electromagnetic wave by looking at the forced oscillations of the electrons in the atoms or molecules of the medium.

Assuming that the electric field intensity of the wave at any given point in the complex representation is of the form

$$\mathbf{E} = \hat{n}E_0e^{-i\omega t}, \quad (14-38a)$$

The polarization at that point is a response of the medium to the periodically oscillating electric field and is of the form

$$\mathbf{P} = \hat{n} \frac{Ne^2}{m} \frac{E_0}{(\omega_0^2 - \omega^2) - i\omega\gamma} e^{-i\omega t}, \quad (14-38b)$$

where e and m stand for the charge and mass of the electron, N denotes the number of electrons per unit volume, γ denotes a damping constant characterizing the motion of an electron, and ω_0 represents the natural frequency of the motion, which is a periodic one by virtue of the fact that the electron is bound in an atom or a molecule where it executes a periodic motion even in the absence of the oscillating field of the electromagnetic wave. The damping constant arises because of the fact that part of the oscillation energy of the electron may get radiated and transferred to the surrounding atoms and molecules.

Equation (14-38b) resembles the expression for the current in a L - C - R circuit receiving supply from an AC source (see eq. (13-55b)) or to the solution of the forced and damped simple harmonic oscillator (eq. (4-52a)) expressed in the complex form (check this last statement out). In all these cases one finds that there is a *phase lag* between a periodic external influence and the steady state response to this influence.

One finds that, in the complex representation, the polarization vector can be expressed in the form

$$\mathbf{P} = \epsilon_0 \tilde{\chi} \mathbf{E}, \quad (14-39)$$

where $\tilde{\chi}$ is termed the *complex susceptibility* of the medium under consideration. The fact that the susceptibility is complex, is indicative of the phase difference between the oscillations of the electric field intensity of the electromagnetic wave under consideration and the oscillations of the resulting dipole moment per unit volume in the medium.

The relative permittivity ϵ_r (also referred to as the *dielectric constant*) of the medium is related to the susceptibility in the form

$$\tilde{\epsilon}_r = 1 + \tilde{\chi}, \quad (14-40)$$

where, like the susceptibility, the relative permittivity is also a complex constant characterizing the medium.

As a consequence of the complex relative permittivity, the medium under consideration

is characterized by a *complex refractive index*

$$\tilde{n} = \sqrt{\tilde{\epsilon}_r}, \quad (14-41)$$

where the relative permeability μ_r of the medium has been assumed to be unity - an assumption which is found to be valid to a good degree of approximation for numerous media of interest.

Finally, the complex wave function for a plane wave propagating along the x-axis is given by

$$\psi = A \exp \left(i\omega \left(\frac{\tilde{n}}{c} x - t \right) \right), \quad (14-42a)$$

while the expression for the wave function with a real value of ϵ and n is given by

$$\psi = A \exp \left(i\omega \left(\frac{n}{c} x - t \right) \right). \quad (14-42b)$$

We now split the complex refractive index into its real and imaginary parts

$$\tilde{n} = n + iq, \quad (14-43)$$

where the real part n and the imaginary part q are, in general, both functions of the angular frequency ω . The dependence on ω can be explicitly worked out from the foregoing equations in this section, but I will skip the derivation in favour of a graphical presentation I am going to give below.

So, the *summary* up to this point is that a plane wave propagating in a dielectric medium induces an oscillating dipole moment in every volume element of the medium by setting the bound electrons of the medium into forced oscillations. The oscillation of the polarization vector differs in phase compared to the oscillation of the electric field vector of the wave under consideration, as a result of which the refractive index becomes complex, with a real part n and an imaginary part q , both of which are, in general, *functions of the angular frequency ω* . This dependence on ω is referred to as *dispersion*.

Substituting eq. (14-43) in (14-42a), one obtains

$$\psi = Ae^{-\frac{\omega q x}{c}} e^{i\omega(\frac{n x}{c} - t)}, \quad (14-44)$$

which, when compared with the expression (14-42b), i.e., the wave function in absence of dispersion, shows that the amplitude of the wave now *decreases exponentially* with x . As a consequence, the intensity of the wave also decreases exponentially: *the wave gets absorbed as it propagates through the medium. Thus, dispersion and absorption go hand in hand.*

The dependence of the real part (n) of the refractive index and of the absorption coefficient ($\alpha \equiv \frac{2q}{\omega c}$; the factor of 2 comes in when one considers the rate at which the intensity gets diminished with distance) on the angular frequency ω derives, in the main, from the right hand side of eq. (14-38b) which, however, needs modification so as to have general applicability. The modification relates to the fact that the electrons in a medium are characterized not by a single natural frequency ω_0 but, in general by a *number of* natural frequencies. In the quantum mechanical picture, the natural frequencies correspond to transition frequencies between bound stationary states of the electrons (see chapter 16 for an outline of a few introductory concepts).

Fig. 14-6 is a schematic graphical presentation of the dependence of the real part (n) of the complex refractive index and of the imaginary part (q) with the frequency ω . One observes that, in general, the refractive index n increases with increasing ω . However, as ω approaches one of the natural frequencies of the medium (the figure shows only one such frequency) the trend of variation is reversed and n *decreases* sharply as ω increases. At the same time, the imaginary part q of the refractive index, and hence the absorption coefficient, which is close to zero away from the natural frequency, increases sharply. The natural frequencies are also referred to as the *resonant* frequencies of the medium (recall the phenomenon of resonance in forced simple harmonic motion (chapter 6) and in AC electrical circuits (chapter 13)). Thus, close to the natural frequencies, the dispersion (i.e., the variation of n with ω) is *anomalous* and, moreover, the medium strongly absorbs the radiation propagating through it. These resonant frequencies are

therefore sometimes referred to as the *absorption edges* of the medium. Away from the absorption edges, the dispersion is *normal*.

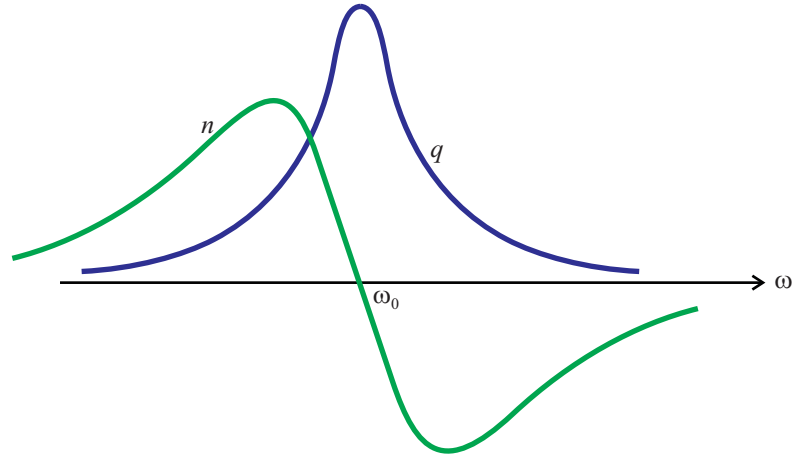


Figure 14-6: Variation of n and q with ω ; only a single resonant frequency (ω_0) is assumed for the bound electrons; close to ω_0 , the trend of increase of the real part (n) of refractive index with ω is reversed (anomalous dispersion); at the same time, q , the imaginary part, which relates to the rate of attenuation of the intensity, increases sharply, being nearly zero for frequencies away from the resonant frequency (absorption edge).

The phase velocity of a monochromatic plane wave of angular frequency ω in the dielectric medium under consideration is given by

$$v_p = \frac{c}{n(\omega)}, \quad (14-45)$$

which, in general, decreases with increasing frequency but increases sharply near an absorption edge. For frequencies just beyond the absorption edge the phase velocity may even be greater than the velocity of light in vacuum.

14.7.1.1 Features of dielectric constant: summary

The above discussion puts into perspective the significance of what was defined as the relative permittivity, or the dielectric constant, of a medium in sec. 11.10.5.2 in the context of static electric fields set up in it. One now finds that the relative permittivity is a function of the angular frequency ω of the electromagnetic field (assumed to vary

harmonically with time) set up in the medium. The constant ϵ_r introduced in the context of static electric fields in the medium is thus just the *limiting value* of this frequency dependent function, corresponding to the limit $\omega \rightarrow 0$ (refer to sec. 11.10.5.6).

Looked at from this point of view, eq. (11-76) is the limiting form of formula (14-40), where the latter underlines the additional feature that the susceptibility and the relative permittivity are *complex* functions, possessing imaginary parts. As we have seen above, this aspect of the relative permittivity is related to the occurrence of *dissipative* processes in the medium as a monochromatic wave propagates through it, which in turn relates to the *phase difference* between the wave and the polarization generated in the medium, the latter being the *response* of the medium to the wave. The response essentially consists of forced oscillations of elementary charges in the dielectric, and *interactions* between a large number of these elementary charges among themselves as also an irreversible loss of energy by radiation from the oscillating electrons.

What is of equal importance relating to the relative permittivity $\tilde{\epsilon}_r(\omega)$ is that it may also have a *field dependence* associated with it, i.e., may depend on the strength of the electric field set up in the medium under consideration. While many of the commonly known media do not exhibit such field dependence of the dielectric constant, a number of other materials, referred to as *nonlinear* ones, do show a such field dependence. Such nonlinear dielectrics have acquired a remarkable importance, from the point of view of applications, in recent decades. However, I will not enter into a discussion on nonlinear dielectric materials in this introductory exposition.

14.7.2 Plane waves in a conducting medium: attenuation

The mechanism of absorption in a dielectric medium involves the transfer of part of the electromagnetic energy of a propagating wave to the bound electrons in the medium under consideration, and the irreversible loss of part of this energy in the form of radiation from the electrons set in forced oscillations and transfer to neighboring atoms and molecules in the medium.

In a conducting medium, on the other hand, energy is lost by a wave propagating

through the medium by way of the *mobile* electrons in the medium gaining energy from the wave and then losing part of this energy irreversibly to the ions in the conductor which are in incessant thermal vibration, and with which the electrons collide in course of their motion produced by the electric field of the wave.

From a mathematical point of view, the description of absorption in a conductor is somewhat similar to that in a dielectric, where in the case of a conductor one finds that the wave-vector k turns out to be a *complex* one, with a real and an imaginary part, where the latter determines the degree of absorption as the wave propagates through the medium. Writing the complex wave-vector as

$$\hat{k} = \hat{i}(k_R + ik_I), \quad (14-46)$$

where the wave is assumed to propagate along the x-axis, its wave function can be expressed in the form

$$\psi = Ae^{-k_I x} e^{i(k_R x - \omega t)}. \quad (14-47)$$

This shows that the amplitude, and hence the intensity of the wave diminishes exponentially as the wave propagates through the conducting medium - the wave gets absorbed, or *attenuated* in the conductor. The phase of the wave at any given point in space is determined by the real part k_R of the wave vector, and the phase velocity is given by

$$v_p = \frac{\omega}{k_R}. \quad (14-48)$$

What is common to the propagation characteristics of a plane wave in a dielectric and a conductor is that, in both the cases, the wave vector (k) is complex. However, in the case of the dielectric, the imaginary part comes in by way of the susceptibility and hence the permittivity being complex (which is, moreover, frequency-dependent). For a conductor, on the other hand, the permittivity is, in general not complex, nor is it frequency dependent to any appreciable degree. Here, the wave vector is complex because of the non-zero *conductivity* (σ) of the material. The real and imaginary parts of k in a

conductor are found to be frequency dependent in a non-trivial way even though the permittivity is not so. This makes the phase velocity, which is determined by the real part of k , frequency dependent, thereby conferring a frequency dependence on the real part of the refractive index.

The degree of attenuation suffered by the wave in the conductor is expressed by the *skin depth*

$$d_{\text{skin}} = \frac{1}{k_I}, \quad (14-49)$$

since the amplitude gets diminished by a large fraction (to nearly $\frac{1}{3}$ of its original value) as the wave propagates through a distance d_{skin} through the medium.

This fact of the wave getting attenuated inside a conductor implies that an electromagnetic wave incident on a conductor fails to penetrate within the bulk of the material since it gets attenuated as it propagates through a distance d_{skin} , thereby remaining confined to a region near the 'skin' of the conductor. At low frequencies, the skin depth decreases with increasing frequency, while at high frequencies the skin depth attains a constant value

$$d_{\text{skin}}^{(\infty)} = \frac{2}{\sigma} \sqrt{\frac{\epsilon}{\mu}}. \quad (14-50)$$

As mentioned above, the phase velocity given by eq. (14-48) also depends on the frequency. At low frequencies, it increases with ω till, at high frequencies it attains the constant value $\frac{c}{\sqrt{\epsilon\mu}}$.

In other words, dispersion and attenuation go hand in hand in a conductor as in a dielectric, though the underlying mechanisms differ in the two media, corresponding to which the features characterizing the propagation and absorption of a wave also differ.

The principal point of distinction relating to wave propagation in dielectrics and in conductors bears repetition here. From the physical point of view, a monochromatic wave propagating through a dielectric sets up forced oscillations of *bound* electrons in the

medium, which are not capable of setting up a current in it. Part of the energy transferred to the bound electrons gets dissipated by means of interactions of the electrons with one another and with the atoms of the medium, as also by irreversible interaction with the surrounding radiation field. By contrast, in the case of a conductor, the electromagnetic field sets up an oscillating current by means of the *free* electrons, where the process of flow of current in a conductor is, in itself, a dissipative process. The dissipation in this case occurs by means of the interaction of the free electrons and the vibrating atoms making up the crystalline structure of the conductor.

14.7.3 Negative refractive index: metamaterials

The refractive index (n) of a medium (determined by its relative permittivity ϵ_r and relative permeability μ_r) is an indicator of how the medium *responds* to the passage of an electromagnetic wave of a given frequency through it. In principle, it should depend on innumerable details relating to the shape, size and mutual arrangement of the atoms constituting the medium. However, since the dimensions of the atoms and their separations are usually small compared to the wavelengths of the electromagnetic waves, the medium effectively acts as a homogeneous one with respect to the wave, and the refractive index depends on just the two parameters ϵ_r and μ_r , all the innumerable parameters at the atomic level being lumped into these two.

For commonly known materials the values of ϵ_r , μ_r and n are all positive (assuming, for the sake of simplicity, that these are all real quantities). However, imagine now that a large number of appropriate microscopic structures (with dimensions small compared to the wavelength) are *implanted* in an otherwise homogeneous medium such that the resulting medium remains, on the whole, an effectively homogeneous one, but the response of the composite system to the passage of electromagnetic waves gets altered, especially in certain wavelength ranges.

The response will now depend on a host of *additional* parameters relating to the nature, size, and arrangement of the implanted units, and will now be characterized by a pair of *altered* parameters (which we continue to denote by ϵ_r and μ_r), and correspondingly a

new refractive index $n(\omega)$, whose dependence on the angular frequency is also likely to get altered.

Such engineered materials are referred to as *metamaterials*. There appears to exist an almost unlimited scope of altering and controlling the bulk properties of materials by such engineering, of which the possibility of artificially generated unusual values of ϵ_r and μ_r is an important and remarkable one. In particular, metamaterials have been fabricated in recent decades for which ϵ_r and μ_r are *both negative* in certain narrow wavelength bands.

For such a material, the propagation of an electromagnetic wave with frequency belonging to the narrow band referred to above, is associated with a number of unusual features. Notable among these is the fact that the group velocity of a wave packet is *oppositely* directed to the phase velocity, and the electric field intensity \mathbf{E} , the magnetic field intensity \mathbf{B} , and the propagation vector of the wave form a *left handed* triad.

What this means is that the *refractive index of the material is effectively negative* (recalling the relation $n = \sqrt{\epsilon_r \mu_r}$, it is theoretically possible to have two signs ('+' and '-') for the square root; for the commonly known materials, the features of wave propagation are consistent with the positive sign while for the metamaterials under consideration, the negative sign is consistent). Among other things, this calls for a re-statement of Snell's law, where both the incident and the refracted ray paths will lie on the same side of the normal to the interface.

Such negative refractive index metamaterials are potentially of great importance from the point of view of applications. For instances, appropriately designed metamaterial slabs can be used as *superlenses* having exceptional imaging and resolution properties.

14.8 The monochromatic spherical and cylindrical waves

Apart from the linearly or elliptically polarized, monochromatic plane progressive wave, other relatively simple solutions of Maxwell's equations (in vacuum or in a material

medium) are also known. Like the plane wave, these solutions, while being idealized ones, are of great importance in electromagnetic theory and optics in that they are found to be relevant in numerous situations of practical interest and, moreover, are of a *basic nature* in that wave disturbances of more general types can be built up by superpositions of these simple wave solutions.

One such set of simple wave solutions are the *spherical waves*. Spherical waves are generated by sources of small size that can, in an approximate sense, be looked upon as *point sources* in much the same way as spherical waves in acoustics are generated by point-like sources (see sec. 9.7). However, electromagnetic waves being of vector nature (recall that the oscillating electric and magnetic field intensities are vectors), the description of these waves is more complicated compared to the scalar spherical waves in acoustics.

Moreover, the description of spherical waves in electromagnetic theory is considerably more involved compared to that of the plane waves as well. There is a certain scheme of classification of these waves where these are identified as *electric multipole* and *magnetic multipole* waves of various *orders*. The simplest and most commonly encountered of these is the *electric dipole* wave. Such a wave is generated, for instance, by a short and straight transmitting antenna in which an alternating current is set up.

The expression for the electric dipole wave of any given frequency, which I do not enter into here, shows that, with the source located at the origin, the magnetic field intensity \mathbf{B} at any point oscillates in a direction perpendicular to \mathbf{r} , but the direction of oscillation of the electric field vector is, in general, *not* perpendicular to \mathbf{r} . In other words, \mathbf{E} , \mathbf{B} , and \mathbf{r} do not form an orthogonal triad. Moreover, the concepts of wave front and wave normal cannot be introduced in a straightforward manner since there is a pronounced *angular dependence* in these spherical waves. In consequence, the direction of energy propagation cannot be related to the wave normal in a simple manner as in the case of the plane wave.

However, at points *far* from the source, the expression for the electric dipole wave as-

sumes a simpler form where it is seen that \mathbf{E} , \mathbf{B} , and \mathbf{r} do form a right handed triad and the wave takes the form of an expanding spherical wave front, with the energy propagating along the wave normal, directed radially.

Finally, at large distances the *intensity* of a spherical wave happens to be a function of $r(= |\mathbf{r}|)$ alone, and varies in proportion to $\frac{1}{r^2}$. This is in contrast to a plane wave where the intensity is constant everywhere in space. However, at such large distances, the spherical wave behaves *locally* as a plane wave, with the electric and magnetic field intensities related to each other as in the case of a plane wave, with the only difference that the amplitude of either varies inversely as the distance r from the source. The inverse square law for intensity follows from this $\frac{1}{r}$ -dependence of the amplitudes, and is consistent with the fact that, at a large distance r , the power radiated from the source is transmitted uniformly through a spherical surface of area $4\pi r^2$.

Analogous to the spherical waves, *cylindrical wave* solutions of Maxwell's equations are also possible, constituting another class of electromagnetic waves of a relatively simple nature. For instance, if a monochromatic plane progressive wave is made to be incident on a long narrow slit whose length is large compared to the wavelength, and if the wave is polarized with its electric field vector oscillating along a direction parallel to the length of the slit, then the wave emerging on the other side of the slit approximates a cylindrical wave (fig. 14-7).

At a large distance from the slit, the wave can be described in terms of a cylindrical wave front with its axis along the slit, where the wave front expands coaxially with a speed c (assuming the wave to be set up in vacuum). The wave is polarized with its electric vector parallel to the length of the slit while the magnetic vector is perpendicular to both the axis and the electric vector. At such large distances from the slit, the intensity of the wave is seen to be *inversely proportional to the distance*. At smaller distances, however, the wave cannot be described in such simple terms.

Analogous to the classification scheme of spherical waves, the cylindrical waves can also

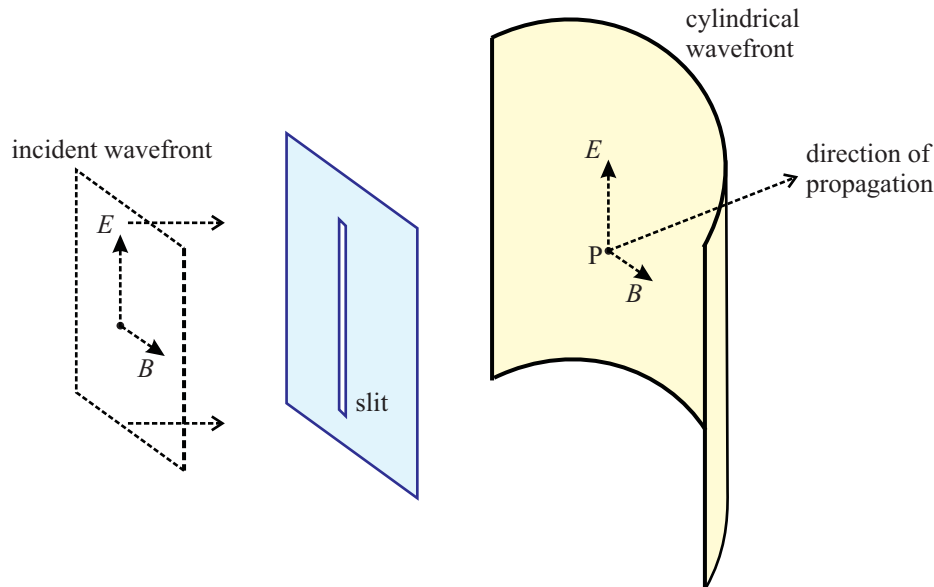


Figure 14-7: Generation of a cylindrical wave; a plane wave front is made to be incident on a long and narrow slit in an opaque screen; the electric field vector of the incident wave is parallel to the length of the slit; the wave emerging on the other side of the slit can be approximately described as a cylindrical wave; far from the slit, it can be described in terms of a wave front which is cylindrical in shape, with the slit as its axis; the wave front expands coaxially away from the slit at a constant rate; the electric vector at any point P is parallel to the axis while the magnetic vector is tangential to the wave front, as shown (dotted arrow); E , B , and the direction of propagation at P form a right-handed triad.

be classified in terms of what can be termed their *multipolarity*, the wave generated as in fig. 14-7 being only one particular instance of these cylindrical waves. A superposition of cylindrical waves of these various types corresponds to a wave disturbance of a more general nature.

14.9 Wave packet and group velocity

As mentioned in sec. 14.3.3, superpositions of simple solutions of the Maxwell equations describing the space-time variations of the electric and magnetic field intensities of an electromagnetic field give rise to solutions of a more complex nature, the latter being often of greater relevance from a practical point of view.

In particular, two or more plane monochromatic waves can be superposed to make up a *wave packet*. This has been explained in 9.15.4 in the context of acoustic waves.

The concept of a wave packet and the principal features described there applies to wave packets made up of plane monochromatic electromagnetic waves too. More generally, one can have wave packets made up of spherical and cylindrical waves as well. While the various different waves making up a wave packet propagate with different phase velocities, the wave packet as a whole propagates with a velocity of its own, termed the *group velocity*. The energy of the electromagnetic field carried by a wave packet propagates through space with the group velocity.

As indicated in sec. 9.15.4, a wave packet is a *coherent* superposition of two or more monochromatic waves (see sections 9.15.1, 14.10, 15.3, 15.3.6.3, and 15.6 for basic ideas relating to coherence), while *incoherent* mixture of a number of monochromatic waves are also of general occurrence.

14.10 Coherent and incoherent waves

In section 14.4 we came across the plane progressive monochromatic wave, described by equations (14-5b), (14-5c). Consider now the variation of the electric and magnetic field intensities given by the following equations,

$$E_y = E_0 \cos(kx - \omega t + \delta) \quad (E_x = E_z = 0), \quad (14-51a)$$

$$B_z = B_0 \cos(kx - \omega t + \delta) \quad (B_x = B_y = 0). \quad (14-51b)$$

where we assume that equations (14-5d) and (14-6) are satisfied, as for the wave described by equations (14-5b), (14-5c), and where δ is a constant which we can choose in the range $0 \leq \delta < 2\pi$. These variations of the electric and magnetic field intensities satisfy Maxwell's equations precisely in the same way as do the variations expressed by equations (14-5b), (14-5c), the only difference being the presence of the constant term δ in the expression for the phase $\Phi' = kx - \omega t + \delta$ as compared to the phase Φ in eq. 14-7a. These equations then correspond to a plane progressive monochromatic wave solution of Maxwell's equations having identical characteristics except for a *phase*

difference δ . In sec. 14.4.8, we came across the idea of a phase difference between two waves while looking at possible superpositions of linearly polarized waves with mutually perpendicular directions of polarization. The two sets of equations, eq. (14-5b), (14-5c), and eq. (14-51a), (14-51b) describe a similar situation where the waves are polarized in the *same* direction.

In reality, a commonly used set-up for the generation of a plane progressive monochromatic wave produces, not a single wave of the form (14-5b), (14-5c), but a large number of waves of the form (14-51a), (14-51b), all with different values of δ . What is more, the values of delta characterizing variations of \mathbf{E} and \mathbf{B} are distributed *randomly* over a set of possible values. A feature of great importance of such an electromagnetic disturbance is the following: if one measures, the electric field intensity \mathbf{E} at any point P at various instants of time t (say, $t = t_1, t = t_2, \dots$), and another observer also measures the variation of the intensity at P at instants of time $t + \tau$ (i.e., at times $t = t_1 + \tau, t = t_2 + \tau, \dots$), where τ is a fixed time-*delay*, then it will be found that, in general, the two variations are not similar to one another or, to put it differently, are not *correlated*. Their resemblance to each other decreases with increasing values of the time-delay τ . One then says that the set-up under consideration produces an *incoherent* wave.

In other words, the wave produced by the given set-up may be described as a superposition of a large number of monochromatic plane progressive waves with randomly distributed phases. At times, such a incoherent wave with random features is referred to as an electromagnetic *disturbance*, to distinguish it from a wave with a well-defined phase, though the more familiar term *wave* is also used.

More generally, the electromagnetic disturbance generated in a given set-up may be looked upon as a superposition of a large number of waves, not only with randomly distributed phases, but with random distributions of other characteristic features like frequency and amplitude as well. For instance, one may have a disturbance made up of waves with frequencies distributed over a narrow range (say, ν to $\nu + \delta\nu$), rather than with those with one single frequency ν .

Still more generally, one may have superpositions of waves with *mutually perpendicular* directions of vibration of the electric (or the magnetic) field intensity vector, and with randomly distributed phase differences. Such disturbances were briefly considered in section 14.4.8 as what are referred to as *unpolarized* plane progressive waves. All these are instances of incoherent waves or, to put it differently, of incoherent electromagnetic disturbances.

There are ways to specify quantitatively the *degree* of coherence of an electromagnetic disturbance, where the latter can range from a *coherent* disturbance (for instance, a plane wave with precisely defined values of amplitude, frequency, and phase, and with a fixed direction of vibration of the electric (or the magnetic) vector), to a completely incoherent one (an unpolarized wave propagating in a given direction, for instance), where disturbances of an *intermediate* degree of coherence are also possible. For instance, consider a mixture of a linearly polarized plane progressive monochromatic wave with an unpolarized wave with the same well-defined frequency and the same direction of propagation. Such a disturbance is referred to as a *partially polarized* wave with features intermediate between those of a polarized coherent wave and an unpolarized, incoherent wave.

Why should a set-up or a source produce a disturbance that is either partially or fully incoherent, while another, more carefully designed, set-up produces a coherent wave?

In general, the electromagnetic disturbance produced by a source depends on very many *internal features* of the source, not all of which can be precisely controlled or specified without fundamentally altering its nature. For instance, the optical radiation emitted by a source of visible light may be the result of an enormously large number of *atomic transitions* where electrons in the atoms making up the source make a jump from one energy level to another inside the atoms (see chapters 16 and 18 for a brief introduction to the energy levels of electrons inside atoms).

The transitions in the various different atoms in the source take place in an essentially uncorrelated manner, independently of one another. Even when the radiation resulting

from these transitions is made to propagate in a given direction, it is highly unlikely that it will have the well-correlated features of a coherent plane progressive monochromatic wave. Similarly, the electromagnetic radiation emitted from a transmitting antenna installed at a TV station is likely to be only partially coherent because the current feeding the antenna may involve fluctuations, depending on the design of the underlying electronic set-up.

A number of phenomena involving electromagnetic waves depend crucially on the coherence properties of the waves, while many others are independent of the coherence features. For instance, coherent and incoherent optical radiations commonly produce very similar visual effects. There are special set-ups however, relating to *interference* experiments in optics (see chapter 15) that do distinguish conspicuously between coherent and incoherent waves.

A considerable number of practical applications involving electromagnetic waves also depend on their coherence properties. Two such applications in optics, of great practical importance, are *holography*, and *optical coherence tomography*, where optical signals with high and low degrees of coherence respectively are made use of.

Ideas relating to coherent and incoherent waves are to be found at several places in this book since these ideas are so important in understanding wave phenomena of various descriptions. Thus, the concept of coherence has been introduced in chapter 9 in the context of interference of acoustic waves. Coherence properties of electromagnetic waves will be taken up again in chapter 15 in connection with *interference and diffraction* of light, and again in the context of the *laser*, the source of coherent light *per excellence*.

14.11 Stationary waves

The idea underlying stationary waves was explained in chapter 9 in the context of acoustic waves in tubes, waves of transverse vibrations in stretched strings, and of vibrations of membranes and plates. In each case, the wave function describing the vibrations at various points in a medium was seen to satisfy a *wave equation* (for the sake of sim-

plicity, only one dimensional and two dimensional wave equations were considered) in a bounded region of space, subject to certain *boundary conditions*.

An alternative and convenient, though less general, way of looking at a stationary wave was also outlined, where a stationary wave was imagined to be set up by the superposition of two traveling waves, one being produced by the reflection of the other at a boundary of the region under consideration.

In the case of electromagnetic waves, the variations of the electric and magnetic field intensities in any given region of space occur in accordance with Maxwell's equations, which imply wave equations involving the field intensities, analogous to the wave equations we encountered for acoustic waves or waves in stretched strings, where now the wave functions satisfying the wave equations are *vector* fields rather than scalar ones (recall, in this context, that elastic waves in solids also involve wave functions of a vector, or more generally, of a tensor nature). As in the case of acoustic waves one is, in general, led to the consideration of three dimensional wave equations for electromagnetic fields which, under special conditions, may reduce to one- or two dimensional wave equations.

If the electromagnetic waves are set up in a limited region of space, so that the electric and magnetic field intensities satisfy certain specific boundary conditions at the surface(s) enclosing the region then the solution to the wave equation assume the form of a stationary electromagnetic wave.

Once again, these standing waves may be interpreted as being the result of superpositions of propagating waves which suffer reflections at the boundary surfaces of the region under consideration.

For instance, consider two propagating electromagnetic waves for which the variations of the electric field intensity are as follows

$$E_y^{(1)} = E_0 \cos(kx - \omega t), \quad E_y^{(2)} = -E_0 \cos(kx + \omega t). \quad (14-52)$$

These represent a pair of progressive waves, of which one (indicated by the superscript '1') propagates along the positive direction of the x-axis and the other (superscript '2') along the negative direction of the x-axis, where both are characterized by the angular frequency ω and wavelength $\lambda = \frac{2\pi}{k}$. The variations of the magnetic field intensities are not shown explicitly here since these are related to the variations of electric field intensities, as explained in sec. 14.4.1.

The negative sign in the expression for $E_y^{(2)}$ is indicative of the relative phase between the two waves, chosen so as to be consistent with the boundary conditions to be specified below (eq. (14-54)).

Let us assume that the above two waves are set up in a limited region of space, given by $0 \leq x \leq L$ so that, in this region the wave resulting from the superposition of the two waves given above can be expressed in the form

$$E = E^{(1)} + E^{(2)} = 2E_0 \sin(kx) \sin(\omega t). \quad (14-53)$$

Let, moreover, the electric field intensity satisfy boundary conditions of the form

$$E = 0 \text{ at } x = 0, x = L, \quad (14-54)$$

at the boundary surfaces ($x = 0, x = L$) of the region. One can then interpret the wave represented by $E^{(2)}$ as the result of reflection of the one represented by $E^{(1)}$ at the surface $x = L$, and conversely, $E^{(1)}$ as resulting from $E^{(2)}$ by reflection at $x = 0$.

The boundary conditions (14-54) then imply that the parameter k determining the wavelength λ (and hence also the angular frequency ω through the relation $\omega = vk$, v being the phase velocity in the medium under consideration) can have any one of the *discrete* set of values given by

$$k = \frac{n\pi}{L}, \quad (n = 1, 2, \dots). \quad (14-55)$$

The stationary wave corresponding to any given value of n is referred to as a *mode* of the electromagnetic field in the bounded region from $x = 0$ to $x = L$. For any such mode, the expression (14-53) shows that the variation of the electric field intensity at any point x is a simple harmonic one, with an amplitude $|E_0 \sin(\frac{n\pi x}{L})|$. The points with maximum amplitude, i.e., the *antinodes*, are the ones with $x = \frac{L}{2n}, \frac{3L}{2n}, \dots$, while those with the minimum amplitude, the *nodes*, are at $x = 0, \frac{L}{n}, \dots$

The variations of the field vectors at all points in the interval within any two successive nodes occur with the *same* phase, while the variations at all points in the next interval occur with the *opposite* phase. The distance between two successive nodes (as also between two successive antinodes) is seen to be $\frac{L}{n} = \frac{\pi}{k} = \frac{\lambda}{2}$.

All these features of a standing wave are analogous to the ones observed for acoustic waves in a tube or for transverse waves in a stretched string. As I have mentioned several times, the wave equation describing the variations of a wave function (the excess pressure, the displacement of a vibrating string, or the electric field intensity of an electromagnetic field) can have various different *types* of solutions, depending on the boundary conditions. Among these, the boundary conditions corresponding to the wave being confined within a finite region of space give rise to standing waves. While I have considered only one dimensional standing waves in this section, more generally one can have standing waves in two or three dimensions as well.

From a basic point of view, the *black body radiation* in an enclosure (see sec. 16.9) can be described as an electromagnetic field where an infinitely large number of standing wave modes are excited, and all these modes are in thermal equilibrium at some given temperature T . Max Planck successfully explained the observed features of black body radiation from a quantum theoretic consideration of these electromagnetic standing wave modes.

Standing waves of large amplitude are set up in a *resonant optical cavity* (see sec. 15.6.7), from which a laser beam of high intensity can be extracted by the operation of an optical shutter.

Chapter 15

Wave Optics

15.1 Introduction

It is light that makes visible all the objects around us, thereby making it possible for us to comprehend the world and the universe at large. That is why a great deal of interest and analysis has been directed for ages to the question of the *nature* of light. Investigations centering around this question have had a long and checkered history. In this book, however, we will not be concerned with the history of these investigations. Instead, I will briefly outline the presently accepted view on the nature of light and, based on this, will tell you how a number of optical phenomena are described and explained.

Light is made up of electromagnetic waves with frequencies lying in a certain range, where this range determines the limits of *visible* light. However, the scope of optics extends even beyond this range. Electromagnetic waves are characterized by frequencies ranging from zero to infinity. Waves of all the frequencies throughout this vast range have a number of general and fundamental properties in common, described and explained in terms of what is known as the *electromagnetic theory* (see chapter 14). In this sense, electromagnetic theory provides the basis of optics. Based on this theory, and making use of a number of *approximation* schemes appropriate for the range of frequencies characterizing light waves, one can explain a large number of optical phenomena. A

number of basic conclusions derived from these approximation schemes constitute the fundamental principles of optics.

Depending on the degree of approximation involved, the approximation schemes mentioned above can be grouped into two subdivisions of optics, namely, *ray optics* (also referred to as ‘geometrical’ optics), and *wave optics* (also termed ‘physical optics’). I have already outlined the basic principles of ray optics in chapter 10, including a number of elementary applications of these principles. In the present chapter I will briefly indicate how wave optics derives from electromagnetic theory and also how ray optics relates to wave optics. In the process I will explain to you a number of optical phenomena from the point of view of the wave nature of optical disturbances, thereby outlining the basic approach of wave optics.

I should mention here, as I have done in the introduction to chapter 10, that electromagnetic theory is not the last word in explaining the nature of light. In the present state of knowledge in physics, the ultimate explanation of the nature of light and of electromagnetic waves in general lies in *quantum theory* (see chapter 16 for an introduction to the principles of quantum theory). Indeed, what I have referred to as wave optics above, derives from this more fundamental quantum theory in a manner similar to the way ray optics derives in turn from wave optics. The approach of describing and explaining various phenomena in optics in terms of quantum theory of radiation is known as *quantum optics*. However, in this book, I will not enter into the principles of quantum optics, instead limiting myself to only a brief reference to the quantum theory of light in outlining the principles underlying the production of coherent light with a *laser* in the present chapter, and in the context of *black body radiation*, *photoelectric emission*, and *Compton scattering* in chapter 16.

The present chapter includes, in the main, an analysis of *interference*, *diffraction*, and *scattering* phenomena in optics. In addition, it includes a brief analysis of the polarization properties of light, which is necessary for a quantitative understanding of various wave phenomena in optics (basic ideas relating to polarization of electromagnetic waves

were introduced in section 14.4.8). I have also included in this chapter a brief introduction to lasers as coherent sources of light, pointing out the quantum concepts involved in an understanding of laser operation. On the other hand, I have not gone into a discussion of *dispersion* in optics. Basic considerations relating to the dispersion of electromagnetic waves in material media have been briefly outlined in sec. 14.7.1.

Much of what I present in this chapter is anticipated and discussed in chapter 9 in the context of acoustic waves. However, while electromagnetic waves (including those in optics) share numerous features of interest with acoustic waves, there are certain crucial points of difference as well, mostly because of the *vector* nature of electromagnetic waves. Additionally, the waves in optics are characterized by extremely small wavelengths that lend them a number of distinctive features.

15.2 Experiments with an illuminated aperture

Fig. 15-1 depicts a plane monochromatic wave incident on a rectangular aperture A in an opaque screen S_1 , there being a screen S_2 on the other side of the aperture. The plane wave corresponds to parallel rays, perpendicular to the wave fronts (see section 14.4.10), incident on the screen S_1 . According to the principles of ray optics, the rays incident on the opaque portions of the screen S_1 would not pass on to the other side, while the rays incident on the aperture would reach the screen S_2 along straight line paths. This should give rise to an illuminated patch on S_2 identical in shape and size to the aperture, while the intensity at all points in the rest of S_2 should be zero (shadow region). In other words, the prediction of ray optics corresponds to a variation of intensity on S_2 as shown in by the rectangular graph in fig. 15-2(A).

In this figure, the x-axis plots the distance along the line $X'OX$ of fig. 15-1 with the origin at O, while the y-axis plots the intensity at various points on this line. One observes that the intensity is constant on a part of the x-axis corresponding to points on the line $X'OX$ falling within the illuminated patch in S_2 , while the intensity falls sharply to zero at points B, B' on the x-axis ($x = \pm \frac{a}{2}$) corresponding to points at the edges of the aperture in S_1 . For all other points on the x-axis, corresponding to the opaque portion

in S_1 , the intensity is zero according to the ray-optics prediction.

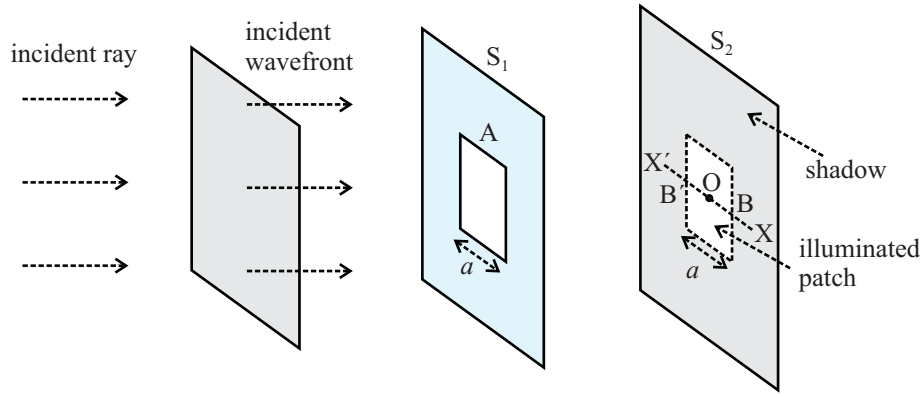


Figure 15-1: Formation of an illuminated patch bordered by a shadow in accordance with the predictions of ray optics (in reality, however, deviations from these predictions are found to occur); a plane wave is incident on the aperture A in the screen S_1 ; this incident wave can be described in terms of a bunch of parallel rays (shown by dotted arrows) incident on S_1 ; if the dimension (a) of the aperture is large compared to the wavelength, a shadow is formed on the screen S_2 with an illuminated patch identical in shape and size to the aperture A ; the intensity distribution on S_2 is shown by the rectangular graph in fig. 15-2(A), where the intensity is plotted against distance along the line $X'OX$.

The observed intensity variation on the screen S_2 does look like the one shown by the rectangular graph in fig. 15-2(A), so long as the dimensions of the aperture A remains *large* compared to the wavelength of light used in the experiment. For instance, for light of wavelength of the order of 500 nm, and an aperture of size of the order of a centimeter, the observed intensity variation is indeed very similar to that corresponding to the rectangular graph.

For apertures of relatively *small* size, however, one observes a deviation from the ray-optics prediction. For instance, with an aperture of size of the order of a millimeter, the intensity variation looks more like the one shown by the curve with the upward and downward swings in fig. 15-2(A).

What is of interest to note in this curve is that it does not correspond to a shadow with a sharp boundary since the intensity at $x = \pm \frac{a}{2}$ does not fall sharply to zero. Instead, the graph swings up and down, the intensity shows an *oscillatory* variation around these

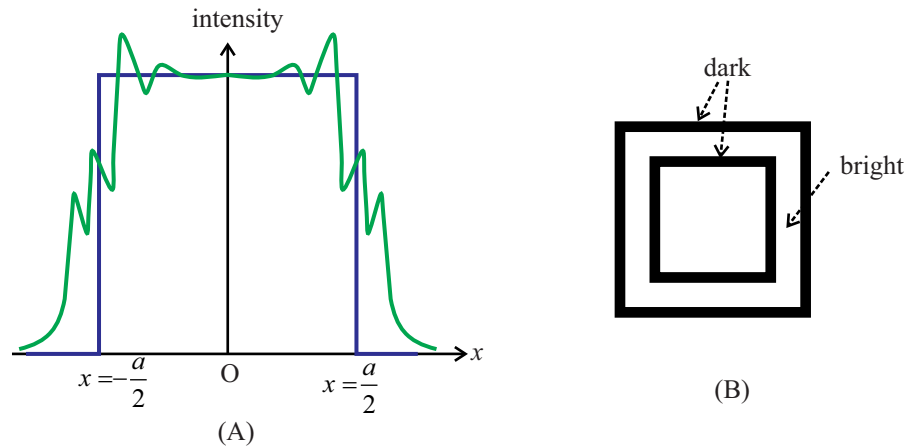


Figure 15-2: Intensity variation on the screen S_2 of fig. 15-1; (A) intensity plotted against distance from O along the line $X'OX$ on the screen S_2 , which cuts the illuminated patch in the segment $B'OB$; the rectangular graph shows the intensity variation for a large aperture where the variation is in accordance with the prediction of ray optics - the intensity being constant within the illuminated patch on the screen and sharply falling to zero at the boundary of the patch; the intensity is zero within the shadow region; the swinging curve shows the nature of the intensity variation for an aperture of small size where the intensity is not zero within the shadow region and, moreover, oscillates near the boundary; (B) illuminated patch on S_2 bordered with alternating bright and dark regions, corresponding to the swinging curve in (A) (schematic).

two points, and there is a non-zero intensity in the shadow region. In other words, the formation of a uniformly illuminated patch in S_2 surrounded by a shadow as predicted by the principles of ray optics does not correspond to the actually observed intensity variation. The latter is characterized by the following features: (a) there is a non-zero intensity in the shadow region, and (b) the border of the shadow region is characterized by oscillations in the intensity, i.e. the formation of bright and relatively dark regions around the shadow border, as shown in fig. 15-2(B).

Fig. 15-3 shows a similar set-up with a bunch of parallel rays (corresponding to a plane wave coming from a distant object) incident on a circular aperture, but now with a converging lens on the other side of the aperture, the observation screen being placed in the focal plane of the lens. The rules of ray optics now predict that the rays coming out of the aperture will be collected by the lens and will be focused at a single point on the observation screen, giving rise to a point image of the distant object. In other words, one expects a bright point on the screen, with the rest of the screen being dark,

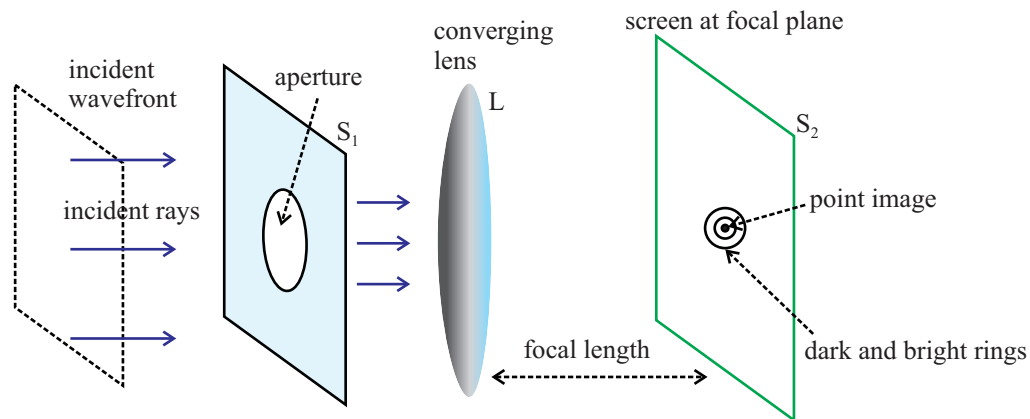


Figure 15-3: Plane wave (corresponding to a parallel bunch of rays originating from a distant object) incident on an aperture, with a converging lens on the other side of the aperture; an observation screen is placed at the focal plane of the lens; according to rules of ray optics, the rays passing through the aperture should be collected by the lens and made to converge at a point on the focal plane, forming the image of the distant object; in reality, one observes alternate dark and bright rings around the image.

corresponding to zero intensity.

In fact, however, one finds a deviation from this prediction of ray optics, the deviation being more pronounced for an aperture of a relatively small size. Close to the image on the observation screen, one finds an intensity variation giving rise to alternate dark and bright rings encircling the image (for an aperture of a different shape, the shape of the dark and bright regions also gets altered). As the aperture is made to be progressively smaller in size, the rings become more pronounced and spread out, covering a larger area on the screen.

15.2.1 Spreading and bending of waves

As I have pointed out earlier at several places in this book, the ray picture is not the last word in optics. A more complete theory looks at visible light as made up of electromagnetic waves in a small frequency range (roughly, 3.8×10^{14} to 1.0×10^{15} Hz). Accordingly, the propagation of light is more appropriately described in terms of the *electromagnetic* theory. This is especially so when the waves encounter apertures or obstacles of size not too large compared to their wavelengths.

As a wave travels past an aperture or an obstacle of such small size, there occurs a *spreading and bending* of the wave, similar to what was indicated in section 9.9. Thus, when a plane wave is incident on a small aperture, the variation of the electric and magnetic field strengths on the other side of the aperture can no longer be described as a plane wave. A schematic representation of the wave disturbance coming out of the aperture looks as in fig. 15-4(A), where the spreading and bending effect near the edge of the aperture is evident, analogous to what one finds in the case of water waves.

For an aperture of relatively large size (~ 1000 times the wavelength), the angle through which the wave is spread and bent is relatively small, while for smaller apertures the spreading covers a wider range of angles. Imagining the radius of a circular aperture to be reduced progressively, one arrives finally at an aperture of size small compared to wavelength of light when the spreading and bending happens to be so large that the wave coming out of the aperture looks like a *spherical wave*, as in fig. 15-4(B).

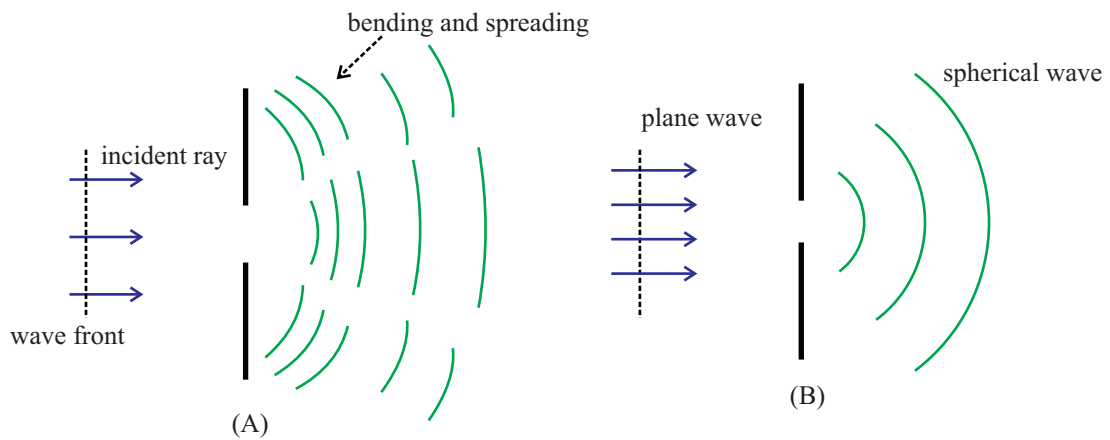


Figure 15-4: Spreading and bending of wave near the edge of an aperture; (A) a plane wave incident on the aperture no longer resembles a plane wave on the other side of the aperture; instead, the wave fronts get bent, corresponding to the wave spreading away from the forward direction; the plane of the figure depicts a section of the aperture and screen shown in fig. 15-1, as also a section of an incident wave front; (B) if the size of a circular aperture is small compared to the wavelength, the bending effect is so pronounced that the variation of the electric and magnetic intensities resembles a *spherical wave*; the plane of the figure depicts sections of a number of spherical wave fronts.

15.2.2 Waves coming out of pin-holes and slits

The fact that a plane wave, incident on a circular pin-hole of dimension small compared to the wavelength, is transformed into a spherical wave on coming out of the pin-hole on the other side, is an interesting result in electromagnetic theory whose derivation I will not go into in this book. The wave coming out of the pin-hole can be described as a spherical wave produced by point-like electric and magnetic dipole sources located at the pin-hole. In other words, from the point of view of describing the wave coming out of the pin-hole, the latter can be looked upon as an *effective point source of radiation*, acting as oscillating electric and magnetic dipoles.

Even these spherical waves from point-like electric and magnetic dipole sources are not as simple objects to describe as a linearly polarized plane progressive wave, because of their rather pronounced angular dependence. However, at a considerable distance from the pin-hole one can describe these waves in terms of a wave front of spherical shape. For a linearly polarized incident wave the spherical wave produced by the pin-hole is characterized by analogous polarization properties, with the electric vector, the magnetic vector, and the direction of propagation (the radially outward direction with the pin-hole at the centre) forming a right-handed triad (fig. 15-5(A)).

Imagine now a linearly polarized plane wave incident on a long narrow slit of width small compared to the wavelength, as in fig. 15-5(B). Once again, the wave spreads out and bends away from the slit, but now only in a direction transverse to its length. In the direction parallel to the length of the slit, there is no bending since the length is large compared to the wavelength and we have seen that bending does not occur for such large dimensions of an aperture. The net result is a *cylindrical* wave (refer to section 14.8) emerging from the slit where, at a sufficiently large distance from the slit, it can be described in terms of wave fronts cylindrical in shape (at smaller distances there is a pronounced angular dependence) with polarization characteristics analogous to a linearly polarized plane progressive wave (fig. 15-4(B)).

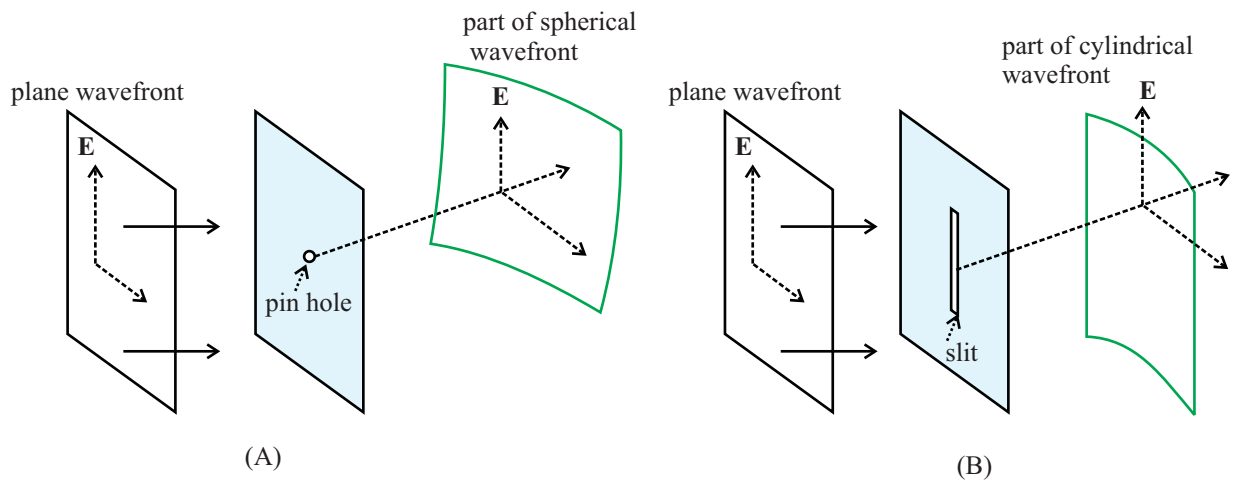


Figure 15-5: A linearly polarized plane wave of optical radiation incident on (A) a pin-hole of circular shape and (B) a long narrow slit; the size of the pin-hole and the width of the slit are small compared to the wavelength; in (A) the wave emerging on the other side of the pin-hole is a spherical wave where a part of a spherical wave front is shown at a considerable distance from the pin-hole; the electric and magnetic vectors, together with the unit vector along the radial direction of propagation form a right-handed triad; in (B) the wave on the other side of the slit is a cylindrical one, with analogous polarization properties.

In summary, there occurs a bending and spreading when a plane wave passes through an aperture which becomes more and more pronounced as the linear dimension of the aperture is made to decrease. If the aperture is reduced to a pin-hole, the wave emerging from it can be described as a spherical wave radiated from a point-like oscillating electric and magnetic dipole located at the pin-hole. At a sufficiently large distance from the pin-hole, the wave can be described in terms of a spherical wave front with polarization characteristics similar to a plane wave. In the case of a slit, a cylindrical wave emerges from it.

This action of pin-holes and narrow slits (with dimensions less than the wavelength of light) will help us understand optical phenomena of *interference* and *diffraction*. The theory underlying the action of pin-holes and narrow slits relates to *scattering* of light, and is a somewhat involved one. Historically, interference and diffraction phenomena were studied and explained before this theory was developed adequately. However, I have brought it up because I feel that it will provide you with a good starting point to develop the study of wave optics from.

1. In reality, the results relating to pin-holes and narrow slits involve a number of simplifying assumptions and limiting procedures. The screen containing the pin-hole or slit cannot be assumed to be made of just any material and of any arbitrary thickness. If one wants to be exact, one has to solve the Maxwell equations subject to appropriate boundary conditions in order to obtain the wave disturbance on the other side of the screen. Indeed, this is what is needed in most *other* problems in optics (or electromagnetic theory) as well. But such solutions are not so easy to come by. The results on pin-holes and slits are related to exact solutions to the Maxwell equations in an ideal and limiting sense. In particular, these are derived by assuming that the screen is infinitely thin and made of perfectly conducting material.
2. There exists a classification scheme of spherical and cylindrical waves in electromagnetic theory (see sec. 14.8) which I will not enter into. The term 'spherical wave' is commonly used in the sense of an electric (or magnetic) dipole wave. Incidentally, in contrast to acoustic waves, spherical electromagnetic waves of the *monopole* type are not possible in principle.

15.3 Interference of coherent waves

15.3.1 Superposition of two plane waves

Consider a situation where two waves of optical radiation pass simultaneously through a given region of space. Assume, for the sake of concreteness, that the electric field vectors of the waves are of the following form

$$\mathbf{E}_1(\mathbf{r}) = \hat{e}E_0 \cos(kx - \omega t), \quad \mathbf{E}_2(\mathbf{r}) = \hat{e}E_0 \cos(kx - \omega t + \delta(\mathbf{r})). \quad (15-1)$$

Both these expressions are similar to that in eq. (14-20a), where each of the waves propagates along the x-axis, with their electric vectors oscillating parallel to \hat{e} , which one may take to be along the y- or the z-axis. What is of significance here is the *phase difference* $\delta(\mathbf{r})$ between the two waves, where this phase difference has been assumed to be a function of the position of the point (\mathbf{r}) under consideration. I have not written out the variations of the magnetic vectors \mathbf{B}_1 and \mathbf{B}_2 here since these can be obtained

by recalling the characteristics of plane electromagnetic waves we are familiar with from chapter 14 and since, moreover, these expressions will not be required now.

Strictly speaking, one needs to check if it is at all possible to have two such waves (as given by the expressions in (15-1)), both satisfying Maxwell's equations with a given set of boundary conditions. Here I have taken a liberty and will have to ask you to *assume* that the second expression satisfies Maxwell's equations just as the first one does though, perhaps, in an approximate sense. In reality, two waves with the same direction of propagation (the x-axis in the present instance) just cannot have a space dependent phase difference between them as I have written above.

However, such a space dependent phase difference can characterize a pair of waves with slightly different directions of propagation. But then, the directions of oscillation of the electric vectors cannot be the same for the two, and will also have to differ. Here I want to tell you what interference in optics essentially consists of, and have allowed for a deviation from Maxwell's equations so as to keep the math at a relatively simple level. The problem with the directions of oscillation of the electric vectors comes up because of the fact that electromagnetic waves involve *vector* wave functions.

As the waves flow through the region under consideration, the variation of resultant electric vector at any point \mathbf{r} is obtained, according to the superposition principle, by adding up the above two expressions. Since the phase difference $\delta(\mathbf{r})$ between the two waves is independent of time, its value at any given point remains constant. Looking at a different point, the phase difference will have a different value but once again it will remain constant in time. In other words, the waves have a definite phase correlation with each other described by the time-independent phase difference $\delta(\mathbf{r})$.

According to what I said in section 14.10, each of the wave disturbances under consideration here is a coherent wave. Moreover, they are seen to have a well-defined and definite value of the phase difference at every point in space in which they get superposed. One then says that the two waves are *coherent* with respect to each other.

The electric field intensity vector for the wave disturbance resulting from the superposition of the two waves (15-1) is

$$\begin{aligned}\mathbf{E} &= \hat{e}E_0 \left(\cos(kx - \omega t) + \cos(kx - \omega t + \delta(\mathbf{r})) \right) \\ &= 2\hat{e}E_0 \cos\left(\frac{\delta(\mathbf{r})}{2}\right) \cos\left(kx - \omega t + \frac{\delta(\mathbf{r})}{2}\right).\end{aligned}\tag{15-2}$$

Comparing with eq. (14-51a), this can be interpreted as a wave propagating along the x-direction with its electric vector oscillating along the unit vector \hat{e} (which we assume to be along the y-axis for the sake of concreteness), the amplitude of the wave at the point \mathbf{r} being $A(\mathbf{r}) = 2E_0 \cos\frac{\delta(\mathbf{r})}{2}$ and its phase being $\Phi(\mathbf{r}) = kx - \omega t + \frac{\delta(\mathbf{r})}{2}$.

The time dependence of \mathbf{E} comes in through the term ωt in Φ and corresponds to a rapid oscillation with frequency $\frac{\omega}{2\pi}$ ($\sim 10^{15}$ Hz), while the space dependence comes in in three ways. The most pronounced space dependence comes in through the term $kx = \frac{2\pi}{\lambda}x$ since the wavelength λ is $\sim 10^{-6}$ m, as a result of which the electric field intensity varies very rapidly with distance along the x-axis. The other two sources of variation are through the phase term $\delta(\mathbf{r})$, one in the amplitude $A(\mathbf{r})$ and the other in the phase Φ . On analyzing the experimental set-up (sections 15.3.4 and 15.3.5) for the production of the waves in (15-1), one finds that these variations are slow compared to the variation by virtue of the term kx in Φ . It is apparent that one can interpret the wave disturbance in (15-2) as a wave with a space-dependent amplitude $A(\mathbf{r})$ and a space-dependent additional phase $\frac{\delta(\mathbf{r})}{2}$.

As noted above, the forms (15-1) of the two waves are, strictly speaking, not in accordance with Maxwell's equations in the context of a realistic experimental set-up. Moreover, the formula (15-2) representing the wave disturbance corresponding to the simultaneous presence of the two waves is also not of strict validity since one needs to ensure that it satisfies the *boundary conditions* for the set-up. In all the instances of interference considered below, the boundary conditions are only approximately satisfied by the superpositions considered in the respective cases.

15.3.1.1 The resultant intensity.

One can then work out the *intensity* of the superposed wave at the point \mathbf{r} . From the result derived in section 14.4.6.2, the intensity is proportional to the modulus squared of the amplitude. In the present instance, then, the intensity at the point \mathbf{r} works out to

$$I(\mathbf{r}) = 4E_0^2 \cos^2\left(\frac{\delta(\mathbf{r})}{2}\right) = 2E_0^2 + 2E_0^2 \cos(\delta(\mathbf{r})), \quad (15-3)$$

where I have set the constant N in eq. (14-13b) at unity since what we will be interested in is not the absolute intensity but a comparison of intensities at various points in space resulting from the superposition of the two waves. This corresponds to setting the *scale* of intensity in such a way that the intensity for a wave of unit amplitude is unity.

Note the slight difference in notation. While E_0 denotes the amplitude of the wave in eq. (14-13b), the amplitude in the present context is space-dependent and is given by $2E_0 \cos(\frac{\delta(\mathbf{r})}{2})$.

The first thing to compare with is the *sum* of the intensities produced at a point \mathbf{r} by the two waves, one independently of the other. Looking at eq. (15-1), each wave corresponds to an intensity $I_0 = E_0^2$ when it passes through the region under consideration all by itself, in absence of the other.

Equation (15-3) then tells us that the *simultaneous* passage of the two waves results in an intensity *different* from the sum of the intensities due to the two waves passing by, one in absence of the other, i.e., from $2I_0$:

$$I = 2I_0 + 2I_0 \cos(\delta(\mathbf{r})). \quad (15-4)$$

15.3.1.2 Maxima and minima in $I(\mathbf{r})$.

Indeed, the intensity is seen to *vary* from point to point in the region under consideration, which contrasts with the fact that the intensity due to a single plane wave is uniform at all points in space. The intensity in expression (15-4) is a maximum ($I_{\max} = 4I_0$)

at those points for which the phase difference $\delta(\mathbf{r})$ is an even multiple of π , while it is a minimum ($I_{\min} = 0$) at those points for which $\delta(\mathbf{r})$ is an odd multiple of π :

$$I_{\max} = 4I_0 \text{ for } \delta(\mathbf{r}) = 0, \pm 2\pi, \pm 4\pi, \dots, \quad I_{\min} = 0 \text{ for } \delta(\mathbf{r}) = \pm \pi, \pm 3\pi, \dots \quad (15-5)$$

What thus happens here is a *redistribution* of energy flux due to the two waves passing through the given region of space, one in the presence of the other. This phenomenon of redistribution of energy flux due to the superposition of two (or more) waves is termed *interference*.

In the above derivation I have assumed for the sake of simplicity that the two waves propagate along the same direction (the x-axis) and are of the same amplitude (E_0). The phenomenon of interference, however, persists even when the waves travel in different directions and are characterized by different amplitudes, provided that the difference in either of these two characteristics is not too large. For instance, if the amplitudes of the two waves are E_1 and E_2 , corresponding to which their intensities are $I_1 = E_1^2$ and $I_2 = E_2^2$, then the maximum and minimum intensities resulting from their interference may be seen to be

$$I_{\max} = (\sqrt{I_1} + \sqrt{I_2})^2, \quad I_{\min} = (\sqrt{I_1} - \sqrt{I_2})^2. \quad (15-6)$$

These expressions reduce to those in eq. (15-5) in the special case considered above, namely when $I_1 = I_2 = I_0$. If the intensities I_1 and I_2 are very much different from one another, say, $I_2 \ll I_1$, one will have $I_{\max} \approx I_{\min}$, i.e., the intensity redistribution will not be appreciable.

In writing eq. (15-6) I have assumed the waves to be propagating in the same direction. For directions of propagation differing considerably from one another the formula for the resultant intensity becomes more involved and, moreover, the redistribution of energy flux becomes less pronounced.

15.3.1.3 Conditions for interference

Two conditions are, however, *necessary* for interference between the waves to occur. The first of these relates to the *frequencies* of the waves. I have assumed the two frequencies to be identical (ω) in the above derivation. This is a necessary condition. If the frequencies are different then the intensity I due to the waves propagating simultaneously will be simply equal to the sum of the individual intensities at all the points in the region under consideration.:

$$I = I_1 + I_2. \quad (15-7)$$

The intensity at a point is the *time-averaged* value of the rate of flow of energy per unit area at that point. For a superposition of two waves of different frequencies, this rate of energy flow per unit area is seen to reduce to zero on averaging over time. However, if each of the two waves is a coherent or an incoherent superposition involving a small range of frequencies, and if the frequency ranges for the two waves overlap, then a redistribution of intensities does occur when the two waves get superposed with each other. What is seen to happen in this case is that the *contrast* between the maximum and minimum intensities becomes less sharp.

The other necessary condition for interference relates to the states of polarization of the waves. Two linearly polarized waves propagating along the same direction do not interfere (i.e., their intensities simply add up as in eq. (15-7)) if their directions of polarization are perpendicular to each other.

However, if the angle between the electric intensity vectors of the two waves differs from $\frac{\pi}{2}$ (or $\frac{3\pi}{2}$), then interference can, in principle, take place.

These results on interference do not involve any new principles other than the principle of superposition, and only require detailed considerations for their derivation. The derivation leading to equations (15-4) and (15-5) was based on a number of simplifying

assumptions since, in this book, the task I have set for myself is to help you get the basic principles right.

15.3.2 A simplified approach: interference of scalar waves

The simple situation considered in the above derivation corresponds to an idealized situation of two waves propagating along the same direction and polarized with their electric (as also magnetic) field vectors parallel to each other.

Recall, however, that a pair of waves of the same frequency, with strictly parallel directions of propagation, cannot at the same time have a space dependent phase difference, because two such waves cannot both satisfy Maxwell's equations with the same boundary conditions.

Once these idealized assumptions are made, one may as well forget about the vector nature of the electric and magnetic field intensities, and consider the simpler problem of superposition of two *scalar* waves, i.e., waves with scalar wave functions. We are familiar with scalar waves from chapter 9 where we found that acoustic waves can be described in terms of variations in excess pressure in a fluid, the latter being a scalar wave function.

Indeed, suppressing the vector nature of the electric field intensity, one can represent the waves in equations (15-1) in the form

$$E_1(\mathbf{r}) = E_0 \cos(kx - \omega t), \quad E_2(\mathbf{r}) = E_0 \cos(kx - \omega t + \delta(\mathbf{r})). \quad (15-8)$$

Considering now the superposition of these two and working out the intensity as the modulus squared of the (space-dependent) amplitude of the resulting field, one arrives at precisely the results given in equations (15-4), (15-5) (check this out).

If the directions of propagation of the two interfering waves make a small angle with each other instead of being along the same line, and if the electric field intensity vectors make a small angle instead of being parallel, then again the simplified derivation using scalar

waves leads to approximately correct results. In this book we will therefore make use of this simplified approach of considering scalar waves. There are instances, however, where the scalar wave approach proves to be inadequate. Such instances require a more elaborate derivation in interference and diffraction.

Situations in which there occurs a redistribution of intensities due to superposition of waves are distinguished from those where there occurs no such redistribution (as in the case of superposition of waves with different frequencies), by referring to the former as resulting in a *summation of amplitudes* and the latter as resulting in a *summation of intensities*.

15.3.3 The complex representation of wave functions

In this section I briefly recall what I presented in section 14.5 in the context of the complex representations of wave functions.

I repeat, first of all, that a common practice in the description of interference and diffraction phenomena in optics is to make use of scalar wave functions in the place of the vector functions \mathbf{E} and \mathbf{B} , a practice that simplifies the mathematics considerably while, at the same time, leading to essentially correct results relating to the intensity distributions in interference and diffraction patterns. One can, if one likes, interpret the scalar wave function as any component of the electric or the magnetic field intensity vector.

1. Considering just one single component of the field vectors does not solve the problem, though, since the other components *cannot* be overlooked. This is because Maxwell's equations introduce an essential interdependence between the various different components.
2. The message, however, remains: calculations for intensity distribution in interference experiments *can* be carried out with a greater attention to details since the basic principles are not different from what I outline here. It is only in the interest of simplicity of the calculations that I consider the superposition of a pair of scalar plane waves propagating along the same direction or, say, a pair of scalar spherical waves. All the while, the waves are assumed to be of the same frequency

and mutually coherent.

Referring to such a scalar wave function, say, ψ , the complex representation makes use of a function $\tilde{\psi}$ where

$$\psi = \text{Re}(\tilde{\psi}). \quad (15-9)$$

In such a complex representation a monochromatic plane wave propagating along the x-axis can be expressed in the form

$$\tilde{\psi} = \tilde{A}e^{i(kx - \omega t)}, \quad (15-10)$$

where \tilde{A} stands for the complex amplitude (refer to section 14.5) and $\Phi = kx - \omega t$ denotes the phase of the wave (at times, one includes only the space-dependent terms in the phase Φ , while displaying the harmonically varying time-dependent part separately).

Monochromatic spherical and cylindrical waves can also be similarly represented in terms of complex wave functions. At times the tilde on top of the symbols representing the sinusoidally varying quantities is suppressed for the sake of convenience, making the intended meanings of the symbols clear whenever there is possibility of confusion.

Set-ups for observing interference.

As we will see in a moment, it is not difficult to produce a pair of coherent waves propagating simultaneously through some region of space by relatively simple optical set-ups where one can observe the spatial variation of the intensity resulting from the superposition of the waves. These are referred to as interference set-ups. The intensity variation from one point to another in the region under consideration produces patterns of alternating bright and dark regions. These are referred to as interference fringes.

What one needs in such a set-up is a *monochromatic source* to ensure that the two waves are characterized by a single common frequency and a pair of pin-holes or narrow slits close to each other for producing the two interfering waves. Modern day *laser* sources constitute high quality monochromatic sources of light. With laser sources, the

separation between the pin-holes or the slits need not be inconveniently small. Other set-ups for the production of interference patterns, like those involving *thin films* are also possible.

15.3.4 Young's pattern with a pair of pin-holes

Fig. 15-6(A) depicts a set-up for the observation of interference fringes produced by the superposition of two coherent optical waves. A coherent plane wave (frequency ν , wavelength λ) is made to be incident on a screen S_1 with a pair of pin-holes Q_1, Q_2 in it, the distance between the holes being, say, d . The pin-holes produce spherical waves on the other side of the screen where there takes place a superposition of the two waves. Interference fringes are formed on an observation screen S_2 placed at a distance D from S_1 , where D is large compared to the wavelength.

For such large values of D , the waves can be described in terms of spherical wave fronts, where the vectors E , B , and the direction of propagation (radial direction with the respective slits at the centres) form a right-handed triad. Considering a region close to the point O (located on the perpendicular bisector of the line joining Q_1 and Q_2), parts of the spherical wave fronts can be looked upon, in an approximate sense, as parts of plane wave fronts (such as P_1, P_2 in fig. 15-6(A)), and one can then employ the results of sections 15.3.1- 15.3.2 to obtain the resultant intensity distribution on the observation screen S_2 .

Though I occasionally refer to a vector wave function to maintain a touch with reality, the derivations are essentially simplified ones arrived at by making a number of assumptions, as I have mentioned above several times. In particular, these derivations apply more appropriately to scalar waves. These are relevant in optics since they give meaningful results in a number of situations of interest. The couplings between the different components of the field variables implied by Maxwell's equations are not of much significance in these situations.

Considering a point P on the screen S_2 in fig. 15-6(A), the waves reaching this point

from Q_1 and Q_2 can be described as in eq. 15-8 where \mathbf{r} is the position vector of P with respect to a right-handed co-ordinate system chosen as follows: the origin is taken at the mid-point of the line joining Q_1 and Q_2 , the x -axis is chosen perpendicular to the planes of the two screens, while the y - and z -axes are chosen parallel to the lines $Y'Y$ and $Z'Z$ respectively.

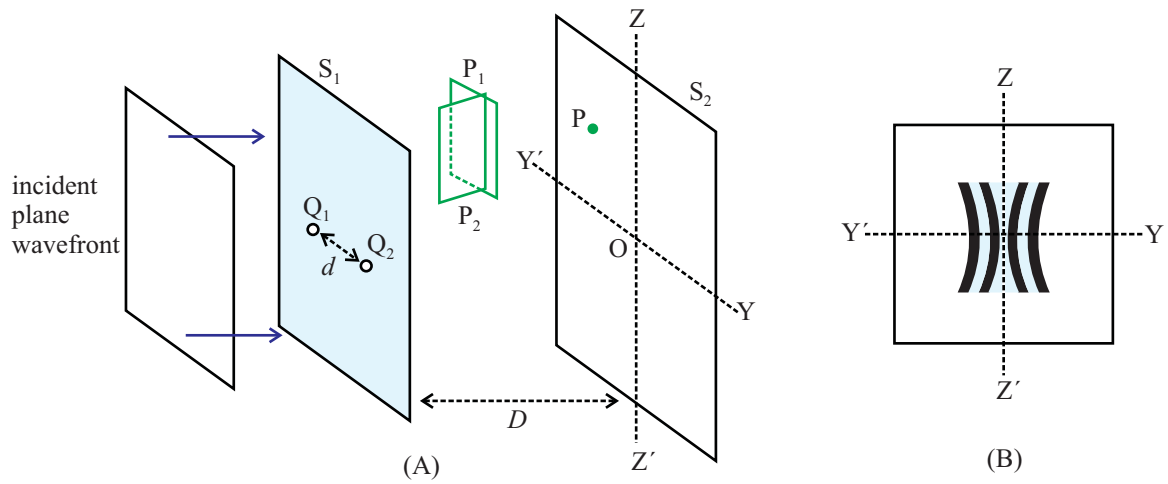


Figure 15-6: (A) A pair of pin-holes Q_1 and Q_2 at a small distance d from each other in an opaque screen S_1 , illuminated with a monochromatic plane progressive wave of light; each of the pin-holes act as a point source of light emitting a spherical wave on the other side of S_1 ; an observation screen S_2 is placed at a distance D from S_1 , where a pattern of interference fringes is observed around the point O ; for a small region around O , the two spherical waves may be looked upon as a pair of coherent plane waves (P_1 , P_2 denote parts of two such wave fronts in a schematic manner) propagating along approximately the same direction; their superposition produces the interference pattern, as can be seen by working out the intensity at any point like P ; (B) the pattern consists of alternating bright and dark hyperbolic fringes.

In order to make use of the results of sections 15.3.1- 15.3.2, one needs to work out the phase difference $\delta(\mathbf{r})$ for any point like P chosen on S_2 . For a plane wave propagating along the x -direction the variation of phase is given by the expression $\Phi = kx - \omega t$ where x stands the distance traversed by the wave front along the direction of propagation. For the point P on the screen S_2 the distances traversed are respectively the lengths of the line segments Q_1P and Q_2P , and hence the phase difference between the two waves is $\delta(\mathbf{r}) = \frac{2\pi}{\lambda}(Q_2P - Q_1P)$. With our choice of the co-ordinate axes, the co-ordinates of the point P , making up its position vector \mathbf{r} are (D, y, z) , while those of Q_1 , Q_2 are, respectively, $(0, -\frac{d}{2}, 0)$ and $(0, \frac{d}{2}, 0)$. Here y and z are the co-ordinates of the point P

with respect to co-ordinate axes $Y'Y$, $Z'Z$ chosen in the plane of the screen S_2 (i.e., ones obtained by translation along the x -axis of corresponding axes chosen on S_1), with O as origin.

With the phase difference calculated as described below, one can put $x = D$ in the expressions in (15-8).

The value of the phase difference $\delta(r)$ is thus given by

$$\delta = \frac{2\pi}{\lambda} \left(\sqrt{(D^2 + (y - \frac{d}{2})^2 + z^2)} - \sqrt{(D^2 + (y + \frac{d}{2})^2 + z^2)} \right). \quad (15-11)$$

For the set-up under consideration, the distance D between the two screens is much larger compared to d , y , and z since the pin-holes are taken to be close to one another and the variation of intensity is observed over a region in S_2 close to O . Ignoring the cubes of these quantities relative to D^3 , one obtains

$$\delta = -\frac{2\pi}{\lambda} \frac{yd}{D} \left(1 - \frac{d^2}{8D^2} - \frac{y^2}{2D^2} - \frac{z^2}{2D^2} \right). \quad (15-12)$$

As a first approximation, one can even ignore $\frac{y^2}{D^2}$, $\frac{d^2}{D^2}$, while retaining $\frac{z^2}{D^2}$ in the above expression, assuming that the region of observation in S_2 extends to a larger distance along the z -axis compared to that along the y -axis. To keep things simple, I will first ignore $\frac{z^2}{D^2}$ and then see how things look when this term is retained.

To a first degree of approximation, then, we have

$$\delta \approx -\frac{2\pi}{\lambda} \frac{yd}{D}. \quad (15-13)$$

Making use of the conditions (15-5), we then see that the intensity alternately assumes maximum and minimum values on either side of the line ZZ' , for the following series of values of y :

$$(\text{maxima :}) \quad y = 0, \pm\lambda\frac{D}{d}, \pm2\lambda\frac{D}{d}, \pm3\lambda\frac{D}{d}, \dots, \quad (15-14a)$$

$$(\text{minima :}) y = \pm\lambda\frac{D}{2d}, \pm\lambda\frac{3D}{2d}, \pm\lambda\frac{5D}{2d}, \dots, \quad (15-14b)$$

In this degree of approximation, then, the fringes are parallel straight lines, the separation between any two successive bright (or dark) fringes being

$$(\text{fringe width :}) w = \frac{\lambda D}{d}. \quad (15-15)$$

In a higher degree of approximation, one needs to retain all the terms in eq. (15-12) (in reality, the terms containing $\frac{d^2}{D^2}$ and $\frac{y^2}{D^2}$ may still be ignored) when the fringes are seen to get bent away from the axis $Z'Z$ as one moves to higher values of z^2 , as shown in fig. 15-6(B). On working out in a consistent manner, the fringes are found to be *hyperbolic* in shape (check this out), being approximately linear (parallel to the z -axis) for small values of z .

This pattern of interference fringes obtained with a pair of pin-hole is termed a *Young's pattern*. One describes the pattern by saying that the two waves produced by the pin-holes interfere *constructively* at the maxima, and *destructively* at the minima. The condition for constructive interference is that the phase difference between the two waves has to be an integral multiple of 2π , while that for destructive interference is that the phase difference has to be an odd integral multiple of π .

15.3.4.1 Phase difference and path difference

In numerous situations of interest, one can alternatively express the above conditions in terms of the *path difference* between the waves. Since the change in phase between two points separated by a distance x along the direction of propagation of the wave is $\frac{2\pi}{\lambda}x$, a phase difference of $2N\pi$ between the two waves corresponds to a path difference of $N\lambda$, while a phase difference of $(2N+1)\pi$ corresponds to a path difference of $(N+\frac{1}{2})\lambda$ ($N = 0, \pm 1, \pm 2, \dots$; the integer N is referred to as the *order* of the fringe under consideration, with maximum or minimum intensity). In other words, a path difference of $N\lambda$ corresponds to constructive interference, while that of $(N+\frac{1}{2})\lambda$ corresponds to destructive interference.

However, in referring to the path difference between the two waves at an observation point (say, P), one has to refer to initial points (say, Q_1 and Q_2) for these and the phase difference is obtained from the path difference by multiplying with $\frac{2\pi}{\lambda}$ only if the phases at these initial points at any given point of time are the same (or differ by $2N\pi$; recall that a phase difference of $2N\pi$ where N is any integer, does not count in the state of a wave at any given point and at any given instant of time). If, for instance, the phases at Q_1 and Q_2 differ by π , then the above conditions for constructive and destructive interference in terms of the path difference will get interchanged.

Optical path.

Finally, the phrase 'path difference' has to be replaced with *optical path difference* in the interest of generality. This is required if the waves under consideration travel in a medium other than vacuum or through more than one such media. Recall the relation $\nu\lambda = v = \frac{c}{n}$ (equivalent to the second and third relations in eq. (14-5d)) where n stands for the refractive index of the medium under consideration. Thus, if a wave traverses a distance x_1 in a medium of refractive index n_1 and a distance x_2 in another medium of refractive index n_2 , then its phase gets changed by $k_1x_1 + k_2x_2$ where k_1, k_2 are given by $k_1 = \frac{2\pi}{\lambda_1} = \frac{2\pi n_1}{\lambda_0}$, and $k_2 = \frac{2\pi}{\lambda_2} = \frac{2\pi n_2}{\lambda_0}$ respectively where the suffixes '1' and '2' refer to the two media, and λ_0 refers to the wavelength of the waves in vacuum (check this out). In other words, the phase change can be expressed as

$$\delta = \frac{2\pi}{\lambda_0}(n_1x_1 + n_2x_2) = \frac{2\pi\nu}{c}\Delta \text{ (say),} \quad (15-16a)$$

where λ_0 stands for the wavelength of the wave in vacuum, ν is the frequency, and

$$\Delta = n_1x_1 + n_2x_2, \quad (15-16b)$$

is the *optical path* traversed by the wave in the two media. This can be generalized in the obvious manner to define the optical path traversed by a wave in more than two media (get this done). For a wave traversing a distance x in a single medium of refractive index n , the optical path is $\Delta = nx$.

Problem 15-1

Imagine two monochromatic point sources of light radiating in phase, located at points with co-ordinates $(0, 0, -d)$ and $(0, 0, 0)$ of a Cartesian co-ordinate system, and an observation point at $(D, 0, 0)$ ($D > 0$). If the wavelength of light from either source is λ , and $d = 15\lambda$, find the largest value of D for which an intensity minimum occurs at the observation point.

Answer to Problem 15-1

HINT: Let S_1 and S_2 be the two point sources and P be the observation point. The path difference of the waves reaching P from S_1 and S_2 is $\Delta = \sqrt{D^2 + d^2} - D$, which is a decreasing function of D (check this out). In order that a destructive interference occurs between the two waves, corresponding to an intensity minimum, one must have $\Delta = (n + \frac{1}{2})\lambda$ ($n = 0, 1, 2, \dots$). The maximum possible value of D , corresponding to the minimum value of Δ satisfying the above condition, is given by $\sqrt{D^2 + d^2} - D = \frac{\lambda}{2}$. Making use of the given value of d , one gets $D = 224.75\lambda$.

Problem 15-2

Referring to fig. 15-6, imagine a small thin transparent plate of thickness $b = 10^{-4}$ m placed against the pin-hole \mathcal{Q}_2 . The wavelength of light used is 600 nm and the refractive index of the plate is $n = 1.6$. Using co-ordinates as in sec. 15.3.4 and considering the observation point with co-ordinates $(D, y, 0)$ ($y > 0$), find the value of y corresponding to the hundredth bright fringe, given that the separation d between the pin-holes is 1.0×10^{-3} m, and the distance of the observation screen from the pin-holes is $D = 0.1$ m.

Answer to Problem 15-2

HINT: As seen in sec. 15.3.4, the phase difference of the waves interfering at the observation point P $(D, y, 0)$, when there is no additional phase difference introduced by means of the thin transparent plate, is, in the first approximation, $-\frac{2\pi yd}{D}$. The thin plate placed against \mathcal{Q}_2 introduces an additional optical path difference nb , i.e., an additional phase difference $\frac{2\pi}{\lambda}nb$. In order that the hundredth bright fringe be formed at P one needs $\frac{2\pi}{\lambda}(-\frac{yd}{D}) + nb = 10 \cdot 2\pi$, which gives $y = 10^{-2}$ m.

15.3.5 Young's pattern with a pair of slits

Figure 15-7 depicts a set-up similar to that in fig. 15-6 with the difference that, instead of a pair of pin-holes, the screen S_1 contains a pair of parallel narrow *slits*, making up what is commonly referred to as a *double-slit* arrangement. With the slits illuminated by a monochromatic plane wave (frequency ν , wavelength λ), each of the slits may now be thought to be producing a cylindrical wave, where the two waves interfere on the other side of S_1 , producing interference fringes that may be observed on S_2 . For sufficiently large values of D , the cylindrical wave fronts look like plane ones, and one may again make use of the results of sections 15.3.1 and 15.3.2 so as to describe the interference fringes at points close to the line ZZ' on S_2 .

We choose a co-ordinate system as in section 15.3.4, with the origin midway between the two slits on S_1 (to be specific, the horizontal line on S_1 may be taken to bisect each of the slits though, in theory, we assume the slits to be infinitely long). The interfering plane wave fronts in this case being approximations to cylindrical wave fronts produced by the slits, the phase for each of the wave fronts will be the same for all points on any line parallel to the cylinder axis, and hence for all points parallel to the axis $Z'Z$ on S_2 .

In other words, the phase difference at any point (y, z) on S_2 between the waves produced by Q_1 and Q_2 will be $\delta = \frac{2\pi}{\lambda}(\rho_2 - \rho_1)$, where ρ_1 , and ρ_2 are distances of the point from the two slits respectively. From the geometry of the arrangement, one then gets

$$\begin{aligned}\delta &= \frac{2\pi}{\lambda} \left(\sqrt{D^2 + \left(y - \frac{d}{2}\right)^2} - \sqrt{D^2 + \left(y + \frac{d}{2}\right)^2} \right) \\ &\approx -\frac{2\pi}{\lambda} \frac{yd}{D}.\end{aligned}\tag{15-17}$$

Recalling, then, the conditions (15-5) for maximum and minimum intensity, one finds that the bright and dark fringes are straight lines given by equations (15-14a) and (15-14b) respectively, there being no bending of the fringes this time for relatively large values of z . The fringe width is given by eq. (15-15). These fringes constitute the *Young's double*

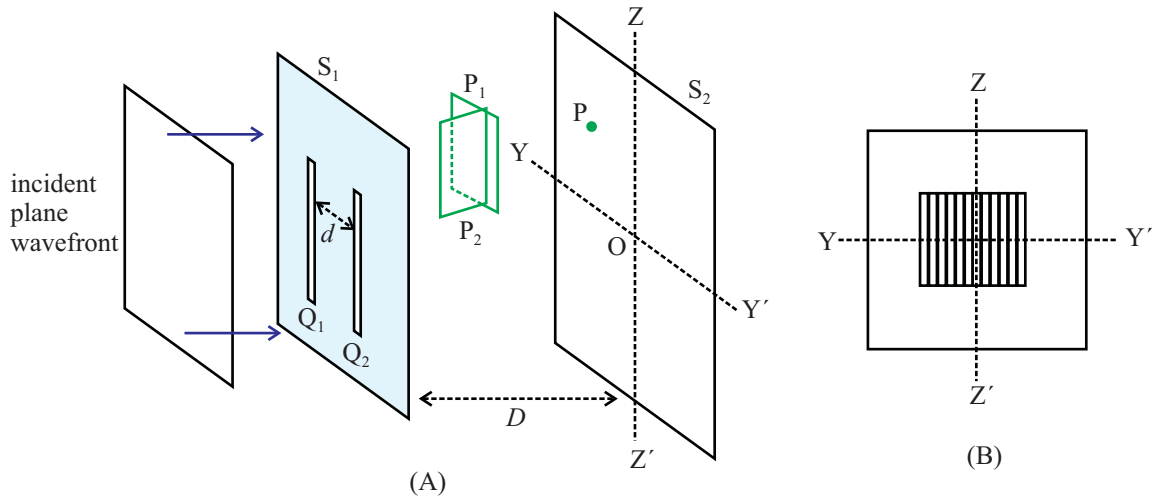


Figure 15-7: A set-up similar to that in fig. 15-6, but now with a pair of long narrow slits Q_1 and Q_2 in place of the pin-holes; the slits produce cylindrical waves on the other side of the screen S_1 , and interference fringes are produced on the second screen S_2 on either side of the line ZZ' ; considering a small region on either side of this line, the cylindrical wave fronts look like plane ones (provided that D is sufficiently large), and interference fringes are formed as the two waves get superposed; (B) the pattern consists of alternating bright and dark straight fringes.

slit pattern.

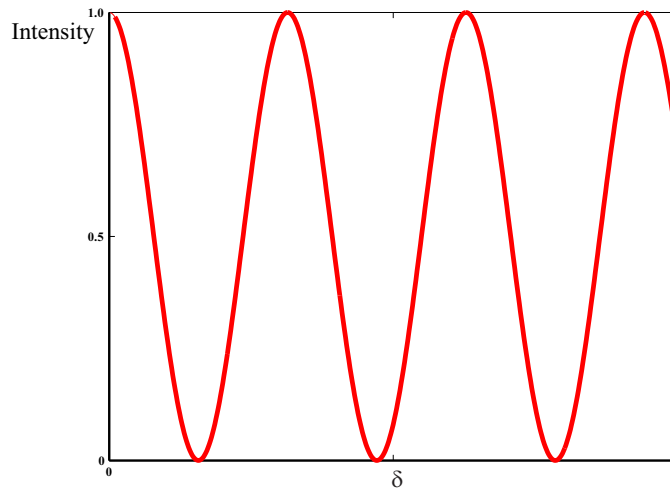


Figure 15-8: Variation of intensity with phase difference δ (on one side of the central maximum) in a double slit interference pattern; the variation of intensity with distance (y) on either side of the central maximum ($\delta = 0$) is of a similar nature; the variation is shown with an arbitrary choice of parameters d and D ; units along the two axes have also been chosen arbitrarily to indicate only the nature of variation; the separation, in terms of δ or y , between successive maxima (or successive minima) of intensity is constant.

Combining equations (15-4) and (15-13) (for the phase difference), the expression for the intensity at a point in the double slit interference pattern is seen to be

$$I = I_0 \cos^2\left(\frac{\delta}{2}\right) = I_0 \cos^2\left(\frac{\pi yd}{\lambda D}\right), \quad (15-18)$$

where I_0 stands for the intensity of a bright fringe (corresponding to $\frac{yd}{D} = n\lambda$, where n is any integer; this differs from the meaning of I_0 in eq. (15-4)), and the meanings of the other symbols have already been explained. In particular, δ stands for the phase difference between the two interfering waves, and y for the distance of the observation point from the central bright fringe ($y = 0$, $\delta = 0$, $I = I_0$). The variation of intensity with delta in a double-hole or double slit interference pattern is depicted in fig. 15-8.

In describing and explaining the double slit interference pattern and the single- and double slit diffraction patterns, I have assumed that the ray paths for the incident plane wave in each case are perpendicular to the screens containing the slits. The analysis can be extended without much difficulty to include situations in which the incident ray paths are inclined at some other angles to the screens.

Problem 15-3

In a double slit interference pattern, if I_0 denotes the maximum intensity, find the angular deviation θ for a point, closest to the central maximum ($\theta = 0$, $I = I_0$), with intensity $I = \frac{I_0}{2}$, in terms of the wavelength λ and the separation d between the slits.

Answer to Problem 15-3

HINT: Referring to eq. (15-17), the angular deviation for a point at a distance y from the central maximum (note from eq. (15-18) that the maxima occur at points for which $\frac{yd}{D} = n\lambda$, where the integer n represents the order of a bright fringe; the central maximum corresponds to $n = 0$) is given by $\theta \approx \frac{y}{D}$, and hence the intensity expression can be written as $I = I_0 \cos^2\left(\frac{\pi}{\lambda}\theta d\right)$. Thus, for $I = \frac{I_0}{2}$, one has $\cos\left(\frac{\pi}{\lambda}\theta d\right) = \pm\sqrt{\frac{1}{2}}$. The minimum magnitude of θ satisfying this condition is $\theta = \frac{\lambda}{4d}$.

15.3.6 Young's fringes with partially coherent light

In analyzing the interference fringes formed by a pair of pin-holes or a pair of slits, I have assumed for the sake of simplicity that the wave incident on the screen S_1 is a monochromatic and linearly polarized plane progressive one, which implies that the wave is completely coherent. Since the pin-holes or the slits introduce a phase difference between the waves which remains constant at every point on the observation screen, the wave disturbance at any observation point continues to be a coherent one, i.e., in other words, one has here a coherent superposition of two coherent waves. One sometimes describes this by saying that the pin-holes or the slits act like a pair of *coherent virtual sources* of light.

In reality, however, it may not be possible to achieve strict coherence for the wave incident on the screen S_1 in the above arrangements. This tends to destroy the interference pattern wherein the intensity distribution has to be described by the *sum-of-intensities* formula rather than the *sum-of-amplitudes* formula, as I mentioned in section 15.3.2.

In the following I consider separately two distinct sources of incoherence of the wave incident on the pin-holes or the slits. The first of these corresponds to the wave being *unpolarized* or partially polarized, while the second corresponds to the wave being not a strictly monochromatic one. In reality, though, unpolarized light can rarely be said to be strictly monochromatic, i.e., it almost always involves a mixture of waves with more than one distinct frequencies.

15.3.6.1 Young's pattern with unpolarized light

An unpolarized plane wave is an incoherent mixture of two linearly polarized plane waves, i.e., a combination of the two waves with no phase correlation between them (in addition to lack of phase correlation, one has to consider lack of amplitude correlation as well; however, I aim to keep the description simple, if a bit incomplete; the conclusions arrived at will be valid, though). For a partially polarized wave, on the other hand, the two linearly polarized waves are characterized by a partial phase correlation. The result of illuminating a pair of pin-holes or slits can be described by referring to the two

linearly polarized components separately. Assuming the directions of linear polarization to be along the z - and the y -axis respectively in figures 15-6 and 15-7, each of the two components produces a fringe pattern independently of the other, each pattern being described by the positions of maxima and minima as in equations (15-14a) and (15-14b).

The intensities due to the two polarized components are now to be *added up* to obtain the resultant intensity distribution due to the unpolarized (or partially polarized) wave because the two components have no phase correlation with each other (sum-of-intensities formula). Since the maxima and minima occur at the *same* positions for the two components, the resultant intensity distribution will also be characterized by maxima and minima at the positions given by equations (15-14a) and (15-14b). In other words, *the use of unpolarized or partially polarized monochromatic light does not affect the Young pattern.*

15.3.6.2 Quasi-monochromatic light

The term *quasi-monochromatic* refers to light (or electromagnetic radiation) made up of monochromatic waves where the frequencies of these monochromatic components lie within a small range, say, from ν to $\nu + \delta\nu$. Since visible light corresponds to frequencies of the order of 10^{14} Hz, even a range as large as $\delta\nu = 10^{10}$ Hz can be considered to be a 'small' one.

Monochromatic light is emitted when an atom makes a transition from one stationary state to another (see chapters 16 and 18). However, a source of light is made up of a large number of atoms and, unless the source is a specially designed one, the transitions of all these atoms occur independently of one another (even though all the transitions occur between identical pairs of states) where, moreover, the atoms undergo a random thermal motion. The *Doppler shift* in frequency (see section 9.13) caused by the motion results in a random spread in the frequencies of radiations emitted from the various different atoms. There are other causes as well for a spread in frequencies to occur, one of which is the *finite lifetime* of the excited state of the atoms from which the transitions take place.

15.3.6.3 Coherence time

The monochromatic components of the radiation emitted from a source are, in some cases, mixed up randomly or incoherently, while in some others there occurs a regular superposition of these components. In general, the variation of any single component of E or B at any given point is no longer a purely sinusoidal one. Rather, it resembles a sinusoidal variation (of frequency $\approx \nu$) for a certain time interval, say, τ , while after a larger interval of time the variations bear no correlation with those within the interval τ . In other words, the correlations are maintained only within a time interval of τ (starting from any given time instant), where the value of τ , referred to as the *coherence time*, depends on the source under consideration.

For a quasi-monochromatic source with a spread $\delta\nu$ in the frequencies of its monochromatic components, the coherence time τ is given by $\tau \approx (\delta\nu)^{-1}$. Now consider the expressions for the phases of two interfering waves at any given point:

$$\Phi_1 = kx - \omega t, \quad \Phi_2 = kx - \omega t + \delta, \quad (15-19)$$

where δ is the phase difference between them. The change in phase of either wave at the given point in a time interval t being ωt , that in an interval τ will be $\omega\tau$. If this be smaller than the phase difference δ between the two waves produced by the virtual sources, then this latter phase difference will be masked by the random fluctuations in phase due to the non-monochromaticity. In other words, the condition for the interference pattern to persist is

$$\omega(\delta\nu)^{-1} > \delta = \frac{2\pi}{\lambda} \frac{yd}{D}, \quad (15-20a)$$

where, in the last expression, I have made use of eq. (15-13) for the phase difference δ between the interfering waves at the observation point (the expression for the phase difference being the same for the slits (eq. (15-17)) as for the pin-holes in the lowest degree of approximation). In other words, with partially coherent quasi-monochromatic light, interference fringes are formed within that region on the screen S_2 for which the

following inequality is satisfied

$$\frac{yd}{D} < c(\delta\nu)^{-1}. \quad (15-20b)$$

A *laser* source of light is characterized by a large value of the coherence time, and a correspondingly small value of the frequency spread $\delta\nu$, and hence with such light, interference fringes are formed up to a much larger distance on either side of the central line $Z'Z$ on the observation screen S_2 in the set-up of fig. 15-6 or 15-7. In other words, a laser source is a highly coherent one and gives rise to interference fringes even when any other source emitting quasi-monochromatic light fails to produce the interference pattern.

A good way of describing a quasimonochromatic disturbance is to represent it as a succession of *wave trains*. Since phase correlations in such a wave at any given point persist up to a time interval τ , it can be represented by an approximately sinusoidal wave within such an interval. This constitutes a 'wave train' of length τ along the time axis, i.e., one of spatial length $c\tau$ (for a wave propagating in vacuum).

As the interval τ is crossed, one has a new interval of duration τ where one finds an approximately sinusoidal and correlated variation, but the phase in this interval bears no correlation with the phase in the earlier interval. In this manner, the wave disturbance at any given point can be looked upon as a succession of wave trains, each of duration τ , successive wave trains being uncorrelated with one another. The propagation of the electromagnetic disturbance then consists of a propagation of this succession of wave trains.

15.3.6.4 Coherence length

Instead of looking at the variations in the electric and magnetic field intensities (or the 'wave functions') as functions of time at a single point in space, one can alternatively look at the variation of a wave function as a function of position at one single instant of time. Since a wave train proceeds with velocity c in free space, the alternative picture, for any given instant of time, is a succession of wave trains in space, each wave train

spanning a distance $L = c\tau$ (fig. 15-9(A)), where L is referred to as the *coherence length* of the wave.

Consider now a partially coherent wave emitted from a source, made up of a succession of wave trains incident on a pair of pin-holes or slits as described above, each of which acts as a virtual source emitting a similar succession of wave trains. At the observation point P, however, the wave trains for the two interfering waves are displaced relative to each other by virtue of the (optical) path difference (Δ) between them. If Δ happens to be larger than L , then the wave trains that get superposed bear no phase relation with each other.

For instance, in fig. 15-9(B) two wave trains A, B for one of the two interfering waves are shown schematically, along with corresponding wave trains A', B' making up the second wave disturbance, for $\Delta > L$. One observes that A' overlaps here with B, but with no part of A. Since A' and A (and similarly B' and B) are corresponding wave trains derived from the incident wave disturbance, the phase in A' is correlated to that in A or to B', but not to the phase in B. In this case, then, an interference pattern will not be formed and the *sum-of-intensities* rule will apply instead of the *sum-of-amplitudes* rule. One observes that the condition $\omega\tau > \delta$ for the formation of an interference pattern is the *same* as the condition $L > \Delta$ arrived at here (check this out).

15.3.7 Thin film patterns

Figure 15-10 shows a monochromatic point source O (which is assumed to be situated in air - or free space - for the sake of simplicity, to be referred to as the *first* medium) and a thin film of a transparent material (of refractive index n , say, the *second* medium) with boundary surfaces AB, A'B'. OP is a ray path incident on the upper boundary surface AB, where the ray path represents the wave normal for a wave front (of frequency ν , say) emitted from O and incident on AB. At the surface AB, the wave suffers reflection and refraction, as a result of which a wave disturbance is sent back into the first medium along the ray path PQ, and another wave disturbance is sent into the second medium along PR. At R, the wave is reflected from the surface A'B' along RS and finally, it is

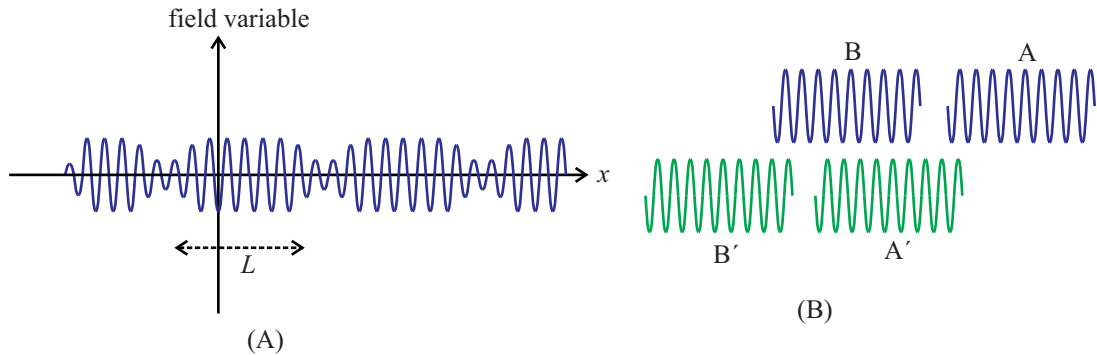


Figure 15-9: (A) Variation of wave function (say, any component of the field intensities) with distance along direction of propagation for a quasi-monochromatic wave; the wave is only partially coherent and is made up of a series of 'wave trains', each spanning a distance $\sim L$, where L is the *coherence length*; the variation within a wave train is similar to a sinusoidal one, but there is no phase correlation between successive wave trains; (B) superposition of two waves, each derived from a partially coherent wave disturbance in an interference set-up; each of the two wave disturbances is made of wave trains (shown displaced in the vertical direction for the sake of clarity), where there is a correspondence between the two sets of wave trains (A with A', and B with B'); however, because of the path difference (Δ), the wave trains are displaced along the x-axis by an amount Δ ; for $\Delta > L$, A' gets superposed with B, with which it has no phase correlation, as a result of which interference does not occur.

refracted at the surface AB along SQ. The wave reflected into the first medium from AB (ray path OPQ and other similar ray paths) and that refracted into the first medium after first suffering a refraction at AB and then a reflection at A'B' (ray path OPRSQ and other similar ray paths) get superposed with each other and result in an intensity distribution determined by their phase difference at various points, essentially in the manner indicated in section 15.3.

For instance, the intensity at the point Q will depend on the phase difference of the waves reaching Q through the ray paths OPQ and OPRSQ. In accordance with what has been said in section 15.3.4.1, this phase difference can be expressed in the form

$$\delta = \frac{2\pi\nu}{c}\Delta + \pi, \quad (15-21a)$$

where the *optical path difference* Δ is given by

$$\Delta = n(\text{PR} + \text{RS}) + (\text{SQ}) - (\text{PQ}). \quad (15-21b)$$

Notice the presence of an *additional term* in eq. (15-21a) corresponding to a phase change of π . A phase change of π occurs when an electromagnetic wave gets reflected from an interface between two media where the wave travels from and gets reflected into the *rarer* of the two media, i.e., the one with the lower value of the refractive index (see fig. 15-11; this is a result that derives from electromagnetic theory, but I shall not get into its derivation in this book).

Thus supposing that the refractive index n of the medium forming the thin layer between the boundary surfaces AB and A'B' is larger than that of the medium above and below the layer (this medium has been assumed to be vacuum or air for the sake of concreteness, but it can be any other transparent medium as well), the wave reflected from the surface AB (at the point P in fig. 15-10) will suffer a phase change of π while that reflected from A'B' suffers no such phase change.

This is why the phase difference between the two waves as given by eq. (15-21a) differs from $\frac{2\pi\nu}{c}$ times the optical path difference. It does not really matter if one writes π or $-\pi$ for the additional phase difference in this equation since one can always add or subtract any multiple of 2π in the phase without altering the value of the electromagnetic field variables, the latter being the physically relevant quantities in the theory (recall that the phase Φ enters into the expression for any of the field variables through a factor of the form $e^{i\Phi}$).

Depending on whether the phase difference δ in eq. (15-21a) is an even or odd integral multiple of π , one may have a maximum or a minimum value of intensity at the point Q relative to points in its neighborhood (check against eq. (15-5)). An intensity distribution is thus brought about throughout the region of space above the thin film, made of a spatial distribution of points with maximum and minimum intensities, making up a system of interference fringes.

Such an interference pattern is said to be produced by a *division of amplitude* as compared with the Young patterns described above, the latter being commonly referred to

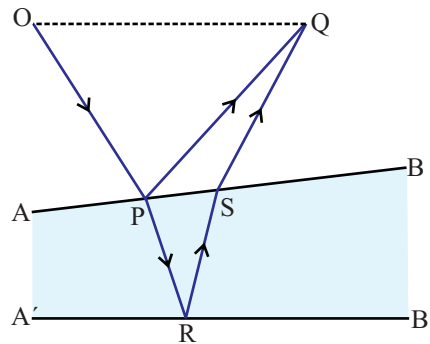


Figure 15-10: Thin film with boundary surfaces AB and A'B'; a wave emitted from the point source O proceeds along the ray path OP and is reflected from AB at P along PQ; part of the wave gets refracted into the second medium (constituting the thin film) along PR and is then reflected from A'B' at R, finally to be refracted at S from AB back into the first medium along ray path SQ; the two waves get superposed in the first medium, and the intensity at any point like Q is determined by the phase difference of the waves reaching the point along the two paths; the direct path OQ is characterized by a large path difference with the above two ray paths and is blocked away.

as interference fringes produced by a *division of wave front*.

1. The use of *ray paths* in describing the formation of fringe patterns requires explanation since interference is a typically wave phenomenon. The concept of ray paths does have a limited relevance in the context of wave propagation, namely, these path indicate the normals to the wave fronts and the paths along which electromagnetic field energy propagates, so long as the wave fronts retain a certain degree of smoothness and regularity of shape. If, however, the wave fronts become disfigured to a considerable extent or get broken up, then the ray paths also lose their relevance.

For a ray path connecting any two given points in space, the phase difference between the wave fronts passing through these two points at any given instant is determined by the optical path length along the ray path as explained in section 15.3.4.1.

For the situation depicted in fig. 15-10 a ray path like OP corresponds to the wave normal of a spherical wave diverging from O, while a ray path like PR corresponds to another wave, of spherical shape, converging to the image point of O (not shown in the figure) resulting from refraction at AB. The optical path difference of two

spherical waves reaching Q along the two ray paths shown in fig. 15-10 is then given by eq. (15-21b).

2. One may wonder as to why the direct ray path from O to Q (dotted line in fig. 15-10) is not to be considered in analyzing the formation of the fringe system. In reality, the optical path difference between the direct wave and any of the waves along the other two ray paths happens to be too large for the direct wave to be relevant in the formation of the fringes. Recall from section 15.3.6.2 that the path difference between the interfering waves has to be less than the coherence length of either of these waves. Even though the source O has been assumed to be monochromatic, it can at best be a quasi-monochromatic one, characterized by a certain coherence length which usually happens to be smaller than the path difference between the direct wave and any of the waves following the other two ray paths shown in fig. 15-10. In an actual set-up the direct wave is blocked off from reaching the observation region so as not to mask the interference fringes.

Since the fringes formed by the thin film fill up an entire region of space, these are referred to as *extended fringes*. Looked at from this point of view, the Young double hole or double slit patterns are also extended fringes, though in the set-ups described in sections 15.3.4 and 15.3.5, only the fringes formed on the observation screen are captured. Similarly, fringes will be found to be formed on a screen held anywhere above the film in the set-up of fig. 15-10. In practice, however, it is difficult to observe these fringes because the set-up involves a *point* source of light, which implies a very low intensity for even the bright fringes in the interference pattern.

This is one reason why a Young pattern formed with a pair of slits is easier to set up and study as compared with a Young pattern produced with a pair of pin-holes. The pin-holes allow too little light to pass through them while the slits, being extended in one direction, produce an interference pattern where the intensity of the bright fringes is much larger and the fringe pattern is distinct.

One then requires an extended source to set up a distinct fringe system with a thin film. Such an extended monochromatic source can be looked upon as an aggregate

of a number of independent point sources like O shown in fig. 15-10. Each of these point sources can be thought to produce a fringe pattern of its own and, the sources being independent of one another, the intensity at any given point like Q is obtained by summing up the intensities due to all these sources acting independently. However, the path difference between the interfering waves at a point like Q due to these point sources will differ from one another, which means that there will be a *distribution* of intensities at Q due to all these sources.

What is more, a similar distribution of intensities will hold for other observation points like Q, and when the intensities due to all the point sources at each observation point is summed up, there results a *uniform* intensity at all the observation points. In other words, *the use of an extended monochromatic source destroys the extended fringe system produced by the thin film.*

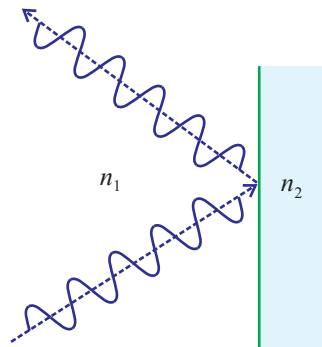


Figure 15-11: Schematic representation of a wave incident on the interface between two media of refractive indices n_1 and n_2 , where the wave, incident from the first medium, is reflected back into it (the wave refracted into the second medium is not shown); if the first medium is optically rarer compared to the second, i.e., $n_1 < n_2$, the wave suffers a phase change of π on reflection; no such phase change takes place for $n_1 > n_2$.

However, while the fringe pattern in the region of space above the thin film disappears, there remain two *special planes on which the fringe pattern persists*. These we now turn to.

15.3.7.1 Fringes of equal inclination

The first of the two special cases we consider relates to a thin film with uniform thickness, i.e., one for which the boundary surfaces (AB and A'B' in fig. 15-10) are parallel to each other, for which fig. 15-12(A) depicts two incident ray paths - the ray path OP originates in the source point O while O'P' originates in another source point O' belonging to an extended monochromatic source. OP gives rise to two ray paths by division of amplitude in the manner depicted in fig. 15-10, but now the ray paths PQ and ST are *parallel* to each other. Similarly, the ray paths P'Q' and S'T' resulting from O'P' are also parallel to each other (reason this out).

In the absence of the converging lens L shown in fig. 15-12(A), the parallel ray paths would meet at infinity where the two waves resulting from a division of amplitude would get superposed. However, the converging action of the lens makes the ray paths PQ and ST meet at some point, say, N (not shown in the figure) in the focal plane of the lens and similarly, P'Q' and S'T' meet at some other point N' (also not in the figure). In other words, a pair of coherent waves get superposed with each other at N and another pair are similarly superposed at N'.

Fig. 15-12(B) depicts the geometry of the ray paths resulting from OP. In this figure, SM is drawn perpendicular to PQ. The optical path difference between the two ray paths is (using symbols shown in the figure)

$$\Delta = n(PR + RS) - PM = 2nt \cos r. \quad (15-22)$$

In other words, the phase difference between the two interfering waves at the point N in the focal plane of the lens L is determined by the thickness (t) of the film, its refractive index (n), and the angle of refraction (r) for the incident ray path at the surface AB.

Problem 15-4

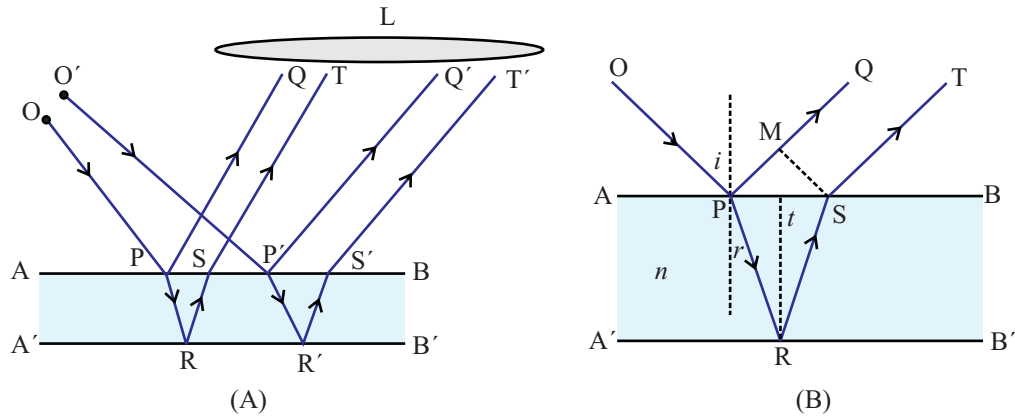


Figure 15-12: (A) ray paths (OP, O'P') originating in two source points (O, O') in the formation of interference fringes by division of amplitude at a thin film of uniform thickness (schematic); OP gives rise to two ray paths PQ and ST, parallel to each other, meeting at the point N in the focal plane (not shown in the figure) of the converging lens L, where the waves corresponding to the two ray paths get superposed; similarly, the ray paths P'Q' and S'T' meet at N' (not shown in the figure) in the focal plane of L, where a pair of coherent waves get superposed; (B) the ray path OP and the two resulting ray paths PQ and ST, where the geometry of the paths determines the phase difference of the waves that get superposed in the focal plane of the converging lens (shown in (A)); i and r are the angles of incidence and refraction at P; SM is drawn perpendicular to PQ; t is the thickness of the film (refractive index n) bounded by the parallel surfaces AB and A'B'.

Check eq. (15-22) out.

Answer to Problem 15-4

HINT: $\Delta = n(PR + RS) - PM = \frac{2nt}{\cos r} - PS \sin i = 2nt \left(\frac{1}{\cos r} - \tan r \sin r \right)$ (invoking law of refraction at P);

i.e., $\Delta = 2nt \frac{1 - \sin^2 r}{\cos r} = 2nt \cos r$.

Considering all incident ray paths making the same angle (i) with the normal to the surface AB (the vertical direction in the figure), the corresponding points of convergence in the focal plane of L will all lie in a circle centered around the focal point (reason this out), and the intensity at all these points will be the same, being determined by the path difference Δ in eq. (15-22), even though the waves reaching these points originate at various different points on the extended source. Put differently, the use of an extended source does not affect the formation of interference fringes in the focal plane of the lens L (or, equivalently, the 'plane at infinity', which is mapped by L into its focal plane). The fringe pattern consists of alternate bright and dark fringes of circular shape, where the

conditions for maximum and minimum intensity are

$$(\text{maxima}) \quad \frac{2\pi\nu}{c}(2nt \cos r) = (2N + 1)\pi, \quad (N = 0, 1, 2, \dots), \quad (15-23a)$$

$$(\text{minima}) \quad \frac{2\pi\nu}{c}(2nt \cos r) = 2N\pi, \quad (N = 0, 1, 2, \dots). \quad (15-23b)$$

1. Check these equations out, making use of equations (15-21a) and (15-22).
2. It is now apparent why one needs a *thin film* to observe these interference fringes: unless the film is thin, the path difference between the interfering waves will be large compared to the coherence length of the waves emitted from the source points.

The circular fringes formed in the focal plane of L are termed *fringes of equal inclination*, since each circular fringe is characterized by a fixed value of the inclination of the various ray paths relative to the normal (vertically upward direction in fig. 15-12) to the film.

15.3.7.2 Fringes of equal thickness

The second of the two special cases mentioned at the end of section 15.3.7 corresponds to a thin film of non-uniform thickness where the plane of observation is *on the film itself*.

Fig. 15-13 shows schematically a wedge-shaped film for which two pairs of incident ray paths are shown, originating in two independent point sources belonging to an extended monochromatic source. The ray paths OP and OQ originate at O while O'P' and O'Q originate at O', where the point Q *lies on the surface of the film* (compare with fig. 15-10). The ray path PRQ produced by refraction at P and a subsequent reflection at R results in the superposition at Q of two coherent waves corresponding to ray paths OQ and OPRQ. Similarly, there occurs a superposition of a pair of coherent waves originating from O' and corresponding to ray paths O'Q and O'P'R'Q.

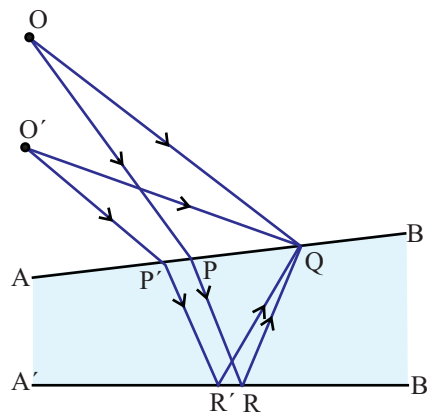


Figure 15-13: Illustrating the principle underlying the formation of fringes of equal thickness; O and O' are point sources belonging to an extended monochromatic source; ray paths OPRQ and O'Q correspond to coherent waves, originating in O, being superposed at Q, where Q lies on the surface of the film; similarly the ray paths O'Q and O'P'R'Q correspond to coherent waves, originating in O', being superposed at the same point Q; the path difference between the first pair of coherent waves does not differ appreciably from that between the second pair.

The path difference for the first pair of coherent waves is determined by the thickness of the film at Q (let us call it t) and the angle of refraction at P, while that between the second pair of waves is determined by the angle of refraction at P', in addition to the thickness of the film at Q. For an extended source of moderate size, and for the observation point Q located on the film, these two path differences turn out to be *close to each other*. Hence, the intensities at P resulting from the interference of all such pairs of waves originating in the different points of the extended source, do not differ appreciably from one another.

Summing up all these intensities, one gets the resultant intensity at Q. Intensities at neighboring points on the film can be similarly worked out and one obtains an appreciable variation of intensity at the various different points on the film, resulting in a fringe pattern. The path difference between the interfering waves at any point like Q depends mainly on the thickness t of the film at that point (regardless of the point of origin of the interfering waves), and thus the intensity variation is determined by the *variation of thickness of the film*.

If one sets up the source in such a way that the incident ray paths are approximately perpendicular to the film (with variations within a small range), then the optical path

difference works out to (see eq. (15-22))

$$\Delta \approx 2nt, \quad (15-24a)$$

and the conditions for maximum and minimum intensity work out to

$$(\text{maxima}) \quad \frac{2\pi\nu}{c} 2nt = (2N + 1)\pi, \quad (N = 0, 1, 2, \dots), \quad (15-24b)$$

$$(\text{minima}) \quad \frac{2\pi\nu}{c} 2nt = 2N\pi, \quad (N = 0, 1, 2, \dots). \quad (15-24c)$$

Since all points on the film with any fixed value of the thickness (t) correspond to the same intensity, the fringe system is made up of alternating contours of maximum and minimum intensity, where each contour traces out points on the film with a constant value of the thickness. These are referred to as *fringes of equal thickness*.

Newton's rings

Fig. 15-14(A) depicts a set-up for the production of fringes of equal thickness. It consists of an extended monochromatic source S, an inclined glass plate P, and a convex lens on a flat reflecting base plate B. Light from the source is partly reflected from the inclined plate P (which is lightly silvered so as to increase its ability to reflect) and is incident more or less normally on the thin air film (plano-concave in appearance) enclosed between the lower surface of the lens and the base plate. Considering the wave sent out from each of the point sources making up S, part of the wave is reflected from the upper surface of the film while another part is reflected from the base plate. The two waves get superposed and form interference fringes of equal thickness in the plane of the short focused telescope (T) conjugate to the surface of the film (effectively the focal plane of the lens combination in the telescope, with respect to which the film is at a large distance; a traveling microscope can be used to make measurements on the fringes).

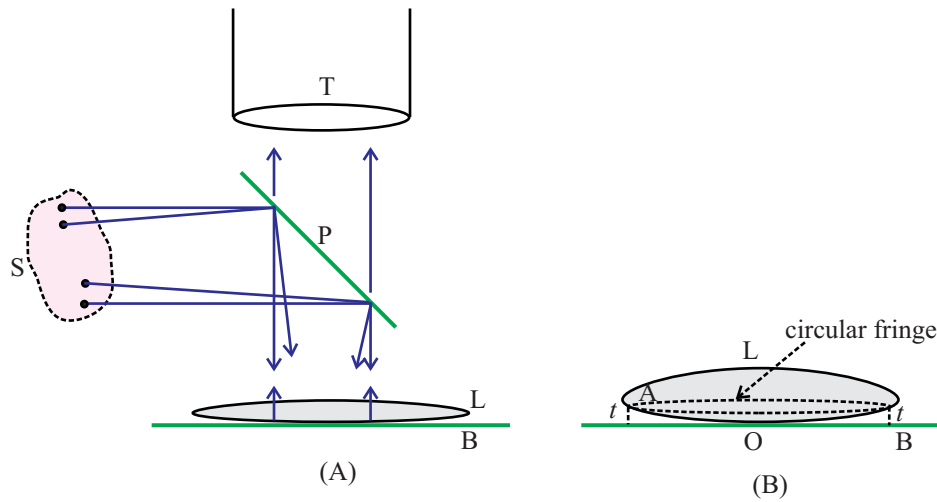


Figure 15-14: Illustrating the formation of Newton's rings; (A) the set-up consisting of an extended monochromatic source S, an inclined glass plate P, and a convex lens L on a plane base plate B; the fringes are viewed through a short focused telescope T focused on the plate; rays from various points on the source get reflected from P (which is lightly silvered so as to increase its ability to reflect) and are incident on the air film between L and B along an approximately normal direction; there occurs a division of amplitude at the surface of the film, and a pair of rays resulting from each incident ray path is sent back towards P; on passing through P, these rays are collected in the telescope T, which forms the image of the fringe pattern developed on the film; the pattern consists of a set of concentric circles, alternately bright and dark; (B) schematic depiction of a circular fringe, formed along the contour of constant thickness t on the plano-concave film between L and B.

As explained below, the fringes are circular in shape for the set-up shown in the figure, and all the source points in S give rise to the *same* set of fringes. In other words, while the fringe system produced by a single source point may not be distinct enough, the fringes produced by the summation of intensities due to all the point sources taken together form a distinct pattern of alternating dark and bright rings, referred to as *Newton's rings*.

Each fixed value of the thickness t corresponds to a circular contour on the film as shown in fig. 15-14(B), where the radius (r) of the contour is related to t as

$$t(2R - t) = r^2. \quad (15-25a)$$

Here R stands for the radius of curvature of the curved surface of the plano-convex film (i.e., of the lower surface of the lens L), which has to be large in order that the film may

be sufficiently thin and the fringe system may be visible - recall that for a thick film the path difference between the interfering waves becomes larger than their coherence length. Using $t \ll R$ in eq. (15-25a), one obtains

$$r^2 \approx 2Rt. \quad (15-25b)$$

Problem 15-5

Check eq. (15-25a) out.

Answer to Problem 15-5

HINT: Referring to fig. 15-14(B), imagine the circle corresponding to the lower surface of the lens L to be completed, and draw the diameter of the circle through the point O, the point of contact with the base plate. Drop a perpendicular from a point on the rim of the dotted circle (point A in the figure) on this diameter, and then make use of the geometry of the circle.

Since any given value of t corresponds to a circular contour of radius r on the film given by eq. (15-25b), the radii of the circular fringes corresponding to maximum and minimum intensities are given by (refer to equations (15-24b), (15-24c))

$$(\text{maxima}) \quad \frac{2\pi\nu}{c} n \frac{r_N^2}{R} = (2N + 1)\pi, \quad (N = 0, 1, 2, \dots), \quad (15-26a)$$

$$(\text{minima}) \quad \frac{2\pi\nu}{c} n \frac{r_N^2}{R} = 2N\pi, \quad (N = 0, 1, 2, \dots). \quad (15-26b)$$

Here n stands for the refractive index of the material of the film enclosed between the base plate and the lower surface of the lens L. While we have taken this to be an air film ($n \approx 1$), it may as well be made of some other material like, say, a liquid of refractive index n . Increasing values of N correspond to circles of progressively larger radius (r_N) and the innermost fringe is a dark one (minimum intensity, $N = 0, r_N = 0$). Counting this as a dark fringe of order zero, the radius r_N of the dark fringe of order N is given by

$$\text{(dark ring)} \quad r_N = \left(\frac{cR}{n\nu} \right)^{\frac{1}{2}} \sqrt{N}. \quad (15-27a)$$

Likewise, the radius of the bright circular fringe of order N is seen to be

$$\text{(bright ring)} \quad r_N = \left(\frac{cR}{n\nu} \right)^{\frac{1}{2}} \sqrt{N + \frac{1}{2}}. \quad (15-27b)$$

Problem 15-6

A broad beam of light of wavelength $\lambda = 500$ nm is made to be incident normally on a thin wedge-shaped film of air formed between two plates inclined at a small angle α to each other. An observer looking at the plates normally observes interference fringes parallel to the edge where the two plates touch. If the distance between the fifth and the forty-fifth bright fringes, counting from the edge of the wedge be $D = 0.4$ mm, find the angle of the wedge.

Answer to Problem 15-6

At a distance d from the edge of the wedge, measured along either of the plates in a direction perpendicular to the edge, the thickness of the air film is αd (check this out), and hence the path difference between the two interfering ray paths is $2\alpha d$ (assuming the refractive index of air to be 1). If this corresponds to a bright fringe, then one has to have $2\alpha d = (N + \frac{1}{2})\lambda$ ($N = 0, 1, 2, \dots$; refer to eq. (15-24b); the wavelength given may be assumed to be that in vacuum ($\lambda = \frac{c}{\nu}$)). Thus, if d_1 and d_2 be the distances corresponding to bright fringes of order $N_1 = 5$ and $N_2 = 45$, then $2\alpha(d_2 - d_1) = (N_2 - N_1)\lambda$, where $d_2 - d_1 = D$. In other words, $\alpha = \frac{(N_2 - N_1)\lambda}{2D} = \frac{40 \times 500 \times 10^{-9}}{2 \times 4 \times 10^{-4}} = 0.025$ radian.

15.3.7.3 The color of thin films

Fig. 15-15(A) depicts a thin transparent film viewed from above in white light, where the light is incident on the film at various different angles, as from a broad source, a number of ray paths coming in from the source at various angles being shown near the top of the figure. In the figure, E denotes the eye of the observer who looks at various

different points on the film such as A, B, C, where it is seen that ray paths reaching the eye from these different points correspond to various different values of the angle of incidence and reflection on the surface of the film.

For instance, the ray path reaching the eye from the point C corresponds to almost normal incidence and reflection while that from the point B corresponds to an angle of incidence (as also of reflection) i on the surface of the film (fig. 15-15(B)).

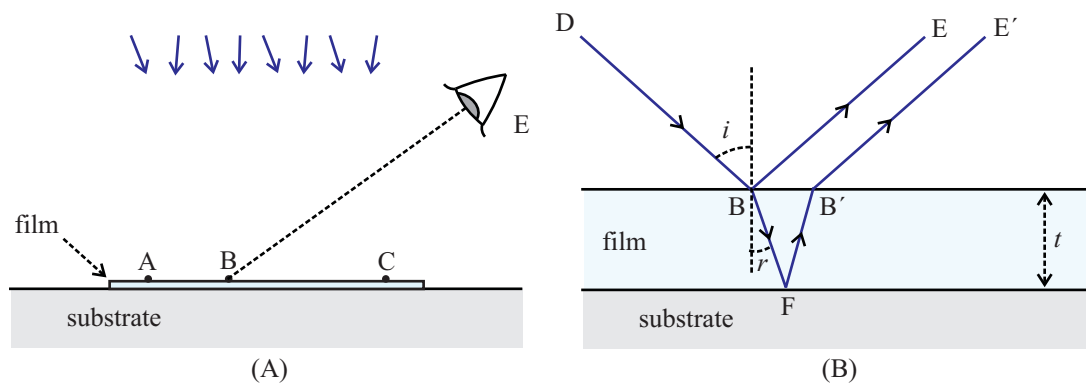


Figure 15-15: (A) Illustrating the origin of the colour developed in a thin film viewed in white light; ray paths from a broad source are incident on the film at various different angles; an observer with her eye at E, on looking at various points like A, B, C on the film receives light along ray paths with various different angles of incidence (i) on the film; (B) explaining the effect observed at any given point B on the film; the effect depends on the phase difference of two waves, one reflected from the upper surface along the ray path BE, and the other reaching the eye along B'E' after being reflected from the lower surface, at the point F (compare fig. 15-12(B), which depicts a similar situation).

Considering any particular direction, say, from B to E, along which a ray path reaches the eye, it may be noted that one has to consider a superposition of *two* waves so as to describe what the eye observes at B. This is explained in fig. 15-15(B) where one finds that an incident ray path (DB) gives rise to two ray paths BE and B'E', one corresponding to a reflection at B on the upper surface of the film while the other corresponding to a reflection from the lower surface (in contact with a second medium commonly referred to as the substrate), analogous to the two ray paths in fig. 15-12(B). Assuming for the sake of simplicity that the two surfaces of the film are parallel to each other, at least over some region around B, the two ray paths BE and B'E' are parallel to each other (and get

superposed in the focal plane of the eye-lens focused for parallel rays), the optical path difference between the two being $\Delta = 2nt \cos r$ (see eq. (15-22)), where n stands for the refractive index of the film and t for the thickness at B.

Assuming for the sake of simplicity that the light is incident on the upper surface of the film from vacuum (or, from air, whose refractive index we assume to be unity), a phase change of π occurs due to the reflection at B (see sec. 15.3.7, referring to eq. (15-21a)). On the other hand, assuming that the substrate is optically denser than the material of the film, a phase change of π occurs at F (on the lower surface of the film in the figure) as well. Thus, in working out the phase difference between the waves corresponding to the two ray paths, these two phase changes cancel each other, and the phase difference is seen to be

$$\delta = \frac{2\pi\nu}{c} 2nt \cos r, \quad (15-28)$$

where the symbols bear the same meaning as in eq. (15-23a).

With the eye in a given position E, the angle r is determined by the point B on the film the eye looks at. Correspondingly, the phase difference δ is determined in terms of ν , the frequency of the light (or equivalently, by the vacuum wavelength $\lambda_0 = \frac{c}{\nu}$). Considering all the various frequency components of the white light incident on the film, those frequencies (say, ν_1, ν_2, \dots ; usually, only one or two of these frequencies belong to the visible part of the spectrum) for which

$$\delta = \frac{2\pi\nu}{c} 2nt \cos r = 2N\pi \quad (N = 1, 2, \dots), \quad (15-29)$$

will correspond to a *maximum* of the intensity observed at B. On the other hand, the frequencies for which

$$\delta = \frac{2\pi\nu}{c} 2nt \cos r = (2N + 1)\pi \quad (N = 0, 1, 2, \dots), \quad (15-30)$$

will be absent in the light reaching the eye from B, since the waves corresponding to the two ray paths interfere destructively for these frequencies.

As a result, the eye will observe a particular coloring at the spot B on the film, and the color will differ for different spots because of the difference in the angle r . This explains the multi-colored appearance of thin transparent films viewed in white light, such as a film of oil on water. The coloring effect gets modified by the absorption of light within the film, since the absorption itself is usually wavelength dependent.

1. In reality, there may be *multiple reflections* at the two surfaces of the film, when one has to consider more than two ray paths in determining the observed effect at any point like B on the film.
2. At times, the film is small in extent, and the variation in the angle r for the various different points on the film is not appreciable, as when the film is viewed, say, perpendicularly. In such cases the entire film appears to be of nearly a uniform color, depending on the wavelength(s) for which a constructive interference takes place.
3. If the substrate happens to be an optically rarer medium compared to the material of the film, then there occurs no phase change of π at the lower surface of the film, and the conditions for constructive and destructive interference get interchanged.
4. For a film of non-uniform thickness, the coloring effect gets modified since now the phase difference of the interfering waves is determined by the thickness of the film at any given point, and the color observed at that point depends on this phase difference.

Problem 15-7

When white light gets reflected normally in air from a soap film of uniform thickness, the reflected light is found to have an interference maximum of intensity at $\lambda_1 = 540 \text{ nm}$, and a minimum at $\lambda_2 = 450 \text{ nm}$, there being no maximum or minimum in between. If the refractive index of the film be $n = 1.35$, what is its thickness?

Answer to Problem 15-7

HINT: For normal incidence, the path difference for the two interfering ray paths is $2nt$, where t denotes the thickness of the film (refer to eq. (15-22)). This corresponds to a minimum intensity if

$2nt = N\lambda$, and to maximum intensity if $2nt = (N + \frac{1}{2})\lambda$ ($N = 1, 2, \dots$; refer to eq.s (15-23a), (15-23b); the wavelengths given are assumed to be those in vacuum ($\lambda = \frac{c}{\nu}$)). Thus, in the present instance, $2nt = (N_1 + \frac{1}{2})\lambda_1 = N_2\lambda_2$. Since there is no interference maximum or minimum in between the two given wavelengths, one has to have $N_1 = N_2 - 1$ (i.e., the maximum possible value of N_1 for a given N_2 ; it is not possible to have $N_1 = N_2$, since $\lambda_1 > \lambda_2$). In other words, $\frac{N_2 - \frac{1}{2}}{N_2} = \frac{\lambda_2}{\lambda_1} = \frac{5}{6}$. This gives $N_2 = 3$, i.e., $t = \frac{N_2\lambda_2}{2n} = 500 \text{ nm}$.

15.3.7.4 Non-reflective coatings

For a film viewed from some particular direction, if the variations in the angles of incidence and reflection (i or r ; refer to sec. 15.3.7.3) for various different directions of vision occur over a small range (say, $i \approx 0$ $r \approx 0$, which holds for a view perpendicular to the film) then the path difference between the two interfering waves produced by reflections at the two surfaces of the film is given by

$$\Delta = 2nt \quad (i \approx 0 \text{ } r \approx 0), \quad (15-31)$$

If the corresponding phase difference equals π , i.e., if

$$t = \frac{c}{4n\nu} = \frac{\lambda_0}{4n} = \frac{\lambda}{4}, \quad (15-32)$$

then the two waves interfere destructively, and the net reflected light reduces to zero. The whole of the light incident on the film is then transmitted through it to the substrate. Here λ_0 stands for the vacuum wavelength of the light incident on the film, while $\lambda = \frac{\lambda_0}{n}$ denotes the wavelength in the material of the film with a refractive index n . In writing eq. (15-31), we have assumed that there occurs an additional phase change of π for both the two interfering waves, i.e., the refractive index of the coating is intermediate between that for air and for the substrate..

Such a ‘quarter wave film’ thus acts as a non-reflecting coating on the substrate. In reality, the light incident on the film may be made up of components with frequencies distributed over a finite range. The film then acts as an effective non-reflecting coating if

the dominant frequency in the combination (i.e., the frequency carrying the largest part of the energy brought in by the optical field) is related to the thickness of the film as in eq. (15-32), and if the energies carried by other frequency components in the range be relatively small.

Non-reflecting coatings are in wide use in lenses and other components of optical instruments as also in various other home and industrial appliances.

While the ‘quarter-wave’ coating makes use of the destructive interference brought about between the two waves reflected from its anterior and the posterior surfaces, and depends for its efficacy on the thickness of the coating being just right so as to effect this destructive interference, another category of non-reflective coatings makes use of a reduced *reflectivity* at the two surfaces by an appropriate choice of the refractive index.

When light is reflected at an interface between two media, the fraction of the incident electromagnetic energy reflected from the interface depends on the relative refractive index of the two media and can be worked out from the principles of wave optics. In the case of a film deposited between air and a substrate, one can effectively reduce the total energy reflected from the two surfaces of the film by appropriately choosing the refractive index (n) of the material of the film relative to that of the substrate (say, n_s), a good choice being $n \approx (n_s)^{\frac{1}{2}}$. Improved performance results if the variation of the refractive index with wavelength happens to be small. Interestingly, the performance of such a coating does not depend appreciably on its thickness.

Often, the above two approaches are appropriately combined in preparing high quality non-reflecting coatings.

15.4 Diffraction of light

15.4.1 Introduction

Let us look back at figures 15-1 and 15-2. How can one account for the intensity variation on the screen S_2 as shown by the swinging curve depicting the variation of

intensity in fig. 15-2(A), the visual appearance on S_2 being as in fig. 15-2(B)? A general observation that has been made is that the incident electromagnetic wave, on passing through the aperture in the screen S_1 , gets bent and fans out away from the forward direction (the direction predicted by ray optics) so that there is some flow of energy into the shadow region.

A similar bending and spreading of the wave was referred to in chapter 9 in the context of acoustic waves. Though an acoustic wave is essentially a scalar wave where the wave function corresponds to the excess pressure while the propagation of light is described in terms of electromagnetic waves where the wave function is made up of a set of vectors, the spreading and bending effects, being essentially wave phenomena, are similar in both the cases. The pattern of intensity variation resulting from the spreading and bending of the wave under consideration is referred to as a *diffraction* pattern, and the phenomenon of spreading and bending itself is termed *diffraction*.

However, just the statement that a wave undergoes a spreading and bending on encountering an obstacle or on passing through an aperture, does not really constitute an explanation of the phenomena, because one has to give a *quantitative* description of the way the wave function (the excess pressure or the electromagnetic field vectors) varies on the other side of the obstacle or the aperture. This enables one to work out quantitative terms the *intensity distribution* in a diffraction pattern, and make comparisons with experimental observations.

In the case of electromagnetic waves, for instance, the ideal thing would be to work out the solution of the Maxwell equations subject to appropriate *boundary conditions*, the latter being a set of constraints that the field vectors have to satisfy at the interfaces of different media like the surface of an obstacle or that of a screen with one or more apertures in it. These constraints ensures that the fields in the various different regions of space (corresponding to media with different properties) are consistent with Maxwell's equations.

However, such a complete solution of the problem is a rare thing in optics. The boundary

conditions in commonly encountered situations prove to be too intractable for working out exact solutions of Maxwell's equations. The few problems for which solutions *can* be exactly worked out correspond to boundary conditions of an idealized nature. For instance, the exact solution of the problem involving a small circular aperture in a flat screen assumes the latter to be an infinitesimally thin sheet of a perfectly conducting material, where a solution in the form of a series can be worked out..

It is in this context that one talks of *diffraction theory*, where the spreading and bending is accounted for quantitatively in terms of a certain approximation scheme. The phenomenon of diffraction is not limited to the field of optics alone. Situations involving electromagnetic waves of frequencies outside the range corresponding to visible light, or ones involving acoustic waves, water waves, seismic waves, or waves of other descriptions, can also be described in terms of diffraction theory.

What, then, is the chief characteristic feature of diffraction? This I state as follows: *the spreading and bending of a wave around an obstacle or aperture whose characteristic linear dimension (d) is large compared to the wavelength (λ), but not too large, constitutes the phenomenon of diffraction.* Typical values of the ratio $\frac{d}{\lambda}$ in diffraction phenomena lie in the range 10^2 to 10^4 where the bending and spreading occurs within moderate limits and the wave fronts do not get altered too much in shape. The behavior of the waves in such situations can be described in terms of an approximate theory which does not lead to exact solutions of the Maxwell equations but which is, in a sense, *consistent* with the latter. Let us now have a very brief introduction to this approximate theory.

15.4.2 The basic approach in diffraction theory

Fig. 15-16 once again depicts an aperture (a rectangular one) in a screen on which a plane monochromatic wave is incident from the left, while P is a point of observation on the other side (to the right) of the screen. The figure shows a section of the rectangular aperture by a plane perpendicular to the screen.

The problem at hand is to quantitatively work out the wave function at P and at similar other points in space.

While the electromagnetic field is described in terms of vector variables (the electric and the magnetic field intensities), the essential nature of the space-time behavior of these variables can be described adequately by working with a *scalar* wave function that can, in an approximate theory, be taken as an appropriately chosen component of the vector field variables. Accordingly, a major trend in the theory of diffraction in optics is to work with a scalar field variable, say, $\psi(\mathbf{r}, t)$, much in the same spirit as in section 15.3.2 in the context of interference. What we will be interested in is the intensity distribution on the ‘shadow side’ of the screen (also referred to as the ‘diffraction region’) where the bending and spreading of the waves is to be accounted for in quantitative terms.

It is not necessarily naive to make use of a scalar wave function in diffraction problems since there exist arguments supporting such an approach - of course, within limits. Technological developments of recent decades have made it necessary that a quantitative theory of diffraction be developed for what is termed ‘wide-angle’ diffraction. For such problems, it becomes necessary to take into account the vector nature of the electromagnetic field variables. However, the theory then becomes mathematically involved to a considerable extent. In this book, I have chosen to confine my presentation to the conceptually and mathematically simpler approach based on scalar wave functions, so as to give you the basic idea of diffraction theory.

A good way to work out the intensity distribution in a diffraction pattern is to employ a *complex representation* of the wave function (refer to section 15.3.3). The complex representation of the wave function for a scalar monochromatic spherical wave of angular frequency ω , for instance, is of the form

$$\tilde{\psi} = \frac{A}{r} e^{ikr} e^{-i\omega t}, \quad (15-33)$$

where r stands for the distance from the point of origin of the wave (the ‘source point’) to the point of observation (the ‘field point’), $\frac{A}{r}$ represents the amplitude of the wave, which varies in inverse proportion to r , and $k = \frac{\omega}{c}$. Here we assume the wave to be set up in free space or in air (in which the velocity of electromagnetic waves is close to c).

The basic idea to start from can now be stated as follows. Due to the bending and spreading of the wave as it moves past the aperture (towards the right in fig. 15-16), the point P *receives signals from all the points in the aperture*. This is in contrast to what the geometrical optics approach tells us, where the point P receives a ray from only one point in the aperture (or none at all, if it happens to fall in the shadow region). The figure shows two points A_1 and A_2 in the aperture from which P receives signals, the latter being in the form of *spherical waves* originating from these two points. In other words, spherical waves originating from all the points in the aperture - referred to as *secondary waves* - may be assumed to move out to any field point like P, where they get superposed to produce the resultant wave function at that point. This, in summary, is the working hypothesis based on which the intensity distribution in a diffraction pattern can be worked out.

1. The idea of secondary waves was made use of by Huyghens in order to explain the propagation of wave fronts. Later developments in the theory in the hands of Fresnel, Kirchoff, and others, showed that the idea of the secondary waves retains a measure of validity even in a more complete theory of diffraction.
2. The question may be asked as to whether secondary waves are *really* emitted from the points on the aperture on which the *primary* wave is incident. Strictly speaking, the only source of light in the set-up of fig. 15-1 or fig. 15-16 is the one located to the far left of either figure from which the plane wave illuminating the aperture (the 'primary wave' as it is called) originates, and there occurs no further emission process from points in the aperture. A convenient and picturesque way to describe the fanning out of the wave in moving past the aperture is to look at the process *as if* the primary wave causes a second set of waves, the secondary ones, to come out from the points on the aperture.

Since the secondary spherical waves from the various points in the aperture all originate due to the same plane wave incident on it, all these are coherent in relation to one another, with their phases being the same at their respective points of origin. Adopting a complex representation of a spherical wave as in eq. (15-33), the spherical waves reaching P from A_1 and A_2 can be represented in the form $\frac{A}{r_1}e^{ikr_1}$ and $\frac{A}{r_2}e^{ikr_2}$ respectively,

where r_1 and r_2 are the length of the ray paths from A_1 and A_2 up to P , and where the time dependence is suppressed for the sake of simplicity (all the respective time dependent factors being identical, namely, $e^{-i\omega t}$), referring only to the complex amplitudes of the waves.

A spherical wave carries electromagnetic energy where the latter propagates along paths perpendicular to the spherical wave fronts, i.e., along radially directed paths, the *ray paths* of geometrical optics.

As all these spherical waves get superposed at a field point like P , the resultant wave function at P is obtained by summing up all the individual wave functions. The variation of each of these wave functions with the position of the point under consideration occurs mainly through the phase factor of the form e^{ikr} , the variation through the amplitude of the form $\frac{A}{r}$ being much slower. This is because of the fact that $k = \frac{2\pi}{\lambda}$, where the wavelength λ is so small that even a very small variation in r leads to a large variation in e^{ikr} .

To a good degree of approximation, then, the wave function of each of the secondary waves can be taken to be of the form Ae^{ikr} where A is a constant which can, moreover, be chosen to be unity while comparing the intensities at various different points in space. The resultant wave function at any given point, caused by the superposition of all these secondary waves, is then obtained by summing up the wave functions coming from all the points like A_1 and A_2 in fig. 15-16.

15.4.3 The intensity distribution

For a sinusoidal time variation of the wave function of the form $\psi(\mathbf{r}, t) = A(\mathbf{r})e^{-i\omega t}$, the intensity at any point \mathbf{r} works out to an expression of the form $I(\mathbf{r}) = N|A(\mathbf{r})|^2$, where N is a constant which we will take to be unity since we will be interested not in the absolute value of the intensity at a point but in the relative intensities at various points in space.

1. In the present context, the space-dependent part $A(\mathbf{r})$ of the wave function is to be obtained by summing up the wave functions coming from all the points in the aperture.
2. The intensity is defined as the time-averaged rate of flow of energy per unit area through a small imaginary surface around the point under consideration, where the surface is assumed to be perpendicular to the direction of energy flow (see sec. 14.4.6.2 for relevant details). The above rule for the calculation of intensity is found to give results in good agreement with experimentally observed intensity distributions and with those worked out from this basic definition of intensity, using more solid reasoning.
3. Let me repeat that I do not enter here into the question of justifying the use of the scalar wave function instead of the vector field variables. For points like P which is not too much away from the forward direction relative to the incident wave, i.e., for points for which the bending and spreading effect is of a moderate degree, this replacement is found to give reasonably good results.

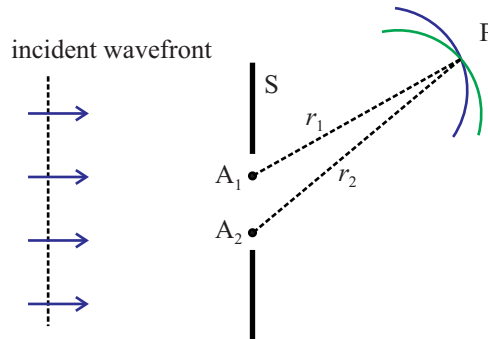


Figure 15-16: Illustrating the basic approach in diffraction theory; a plane monochromatic wave incident on an aperture in the screen S is diffracted into the region on the other side of the screen; the wave function at any given point P in this region can be looked upon as being a superposition of those corresponding to secondary waves emitted from all points in the aperture, two such points and the corresponding secondary waves being shown; each secondary wave is associated with a ray path and a corresponding path length (r_1 and r_2 corresponding to the secondary waves from points A_1 and A_2).

Since the aperture consists of a continuously distributed set of points, one has to imagine it to be partitioned into a large number of area elements, with the area of a typical element being, say, δs . Each such element can then be imagined to be the source of a

secondary wave, whose amplitude is proportional to δs . One thus has to sum up contributions of the form $e^{ikr} \delta s$ from the various area elements. In the limit of the area δs of each element going to zero, this sum reduces to an integral over the entire area of the aperture.

Finally, then, one arrives at the following approximate expression for the relative intensity at any given point P in the diffraction region

$$I = \left| \int e^{ikr} ds \right|^2, \quad (15-34)$$

On working out this expression, one finds that the intensity varies from point to point, with alternating sets of points of maximum and minimum intensity as in fig. 15-2. These alternating bright and dark sets of points are referred to as *diffraction fringes*.

The above outline of the basic theory of diffraction is based on a number of approximations and idealized assumptions, and is not a foolproof or logically sound theory. As I have already mentioned, the only sound theory has to be one that works out the solution to Maxwell's equations, subject to appropriate boundary conditions. In the absence of a well-charted procedure for arriving at such solutions, approximation schemes are to be followed, one such scheme being the approach outlined above. Strictly speaking, the concept of secondary waves is nothing more than a convenient means for arriving at meaningful results relating to diffraction set-ups. While more sophisticated and improved theories of diffraction exist, these have one thing in common with the simplified theory outlined above: none of these involves an actual solution of Maxwell's equations with realistic boundary conditions, except in a few rare instances.

On looking at eq. (15-34), it is seen to possess a simple structure: *the relative intensity is the squared modulus of a sum (or integral) of terms of the form e^{ikr} , where r stands for the length of the ray path from a typical point in the aperture to the field point under consideration, the ray path being the path along which electromagnetic energy carried by*

the secondary wave reaches the field point.

The formula, however, needs to be modified in one important respect. Comparing, for instance, fig. 15-1 and 15-3, one finds that, in the latter, the secondary waves reach the point of observation only after passing through the converging lens placed to the right of the aperture, while in the former, the secondary waves reach the field point directly. Accordingly, the ray path along which the energy carried by the secondary wave propagates may be made up of segments, where the different segments correspond to media of various different refractive indices. As a result, the path length r is to be replaced with the *optical path length* (refer to section 15.3.4.1) along the ray path under consideration.

15.4.4 Fraunhofer and Fresnel diffraction patterns

Figures 15-1 and 15-3 depict diffraction set-ups of two different kinds. In the former, the bending and spreading of an incident plane monochromatic wave past an aperture causes a *loss of sharpness of the boundary between the illuminated patch on the observation screen and the shadow region surrounding it*, with the appearance of a fringe pattern made up of intensity maxima and minima in this boundary region. In the latter, on the other hand, ray optics predicts the formation of a sharp illuminated point in the focal plane of the lens, this illuminated point being the image of the point object that may be assumed to be located at an infinitely large distance from the aperture, sending out a bunch of parallel rays incident on it. The bending and spreading of the wave past the aperture then causes a *fringe pattern to appear around the image*. The pattern again consists of alternate bright and dark fringes, this time blurring the definition of the image point.

These two types of diffraction phenomena are referred to as *Fresnel* and *Fraunhofer* diffraction respectively. Among these, Fraunhofer diffraction may be considered to be a special case of Fresnel diffraction. Consider, for instance, the set-up of fig. 15-3, with the modification that now the observation screen is placed either in front of (fig. 15-17(A)) or behind (fig. 15-17(B)) the focal plane of the converging lens. The ray paths

predicted by ray-optics then result in an illuminated patch on the observation screen instead of a single bright image point. Once again, the bending and spreading of the wave in moving past the aperture results in a loss of sharpness of the boundary between the illuminated patch and the surrounding shadow region, and there appear diffraction fringes on either side of the boundary, similar to those observed in the set-up of fig. 15-1. These are the ones referred to as the Fresnel diffraction fringes. Unless otherwise stated the sources of light in Fresnel and Fraunhofer diffraction set-ups will be assumed to be monochromatic.

In other words, the set-up of fig. 15-3 in general produces Fresnel diffraction fringes *except when* the observation screen is placed in the focal plane of the lens. In the special case of the observation screen being in the focal plane of the lens, the illuminated patch reduces to the point image in the ray-optics description, and the fringes around this point image, produced by the bending and spreading effect, are the ones referred to as the Fraunhofer diffraction fringes.

This fact of the Fraunhofer pattern being a special case of the Fresnel pattern may also be illustrated with the help of the set-up of fig. 15-1 where, in general, a Fresnel pattern is observed for any arbitrarily chosen position of the observation screen (S_2). However, in the special case when the latter is placed at an *infinitely large* distance from the aperture, the diffraction pattern on it assumes a distinctive form since now the illuminated patch on the screen, formed in accordance with the rules of ray optics, reduces to a point, and the intensity distribution around the point resembles the Fraunhofer pattern of fig. 15-3. Indeed, the lens in fig. 15-3 serves the purpose of transforming the plane at infinity of fig. 15-1 to the focal plane where the geometrical image of the object point (in the present instance, a point at an infinitely large distance to the left of the aperture) is formed.

In summary, then, one can describe a Fresnel diffraction pattern as a *fringed shadow*, while an Fraunhofer diffraction pattern can be described as a *fringed image*.

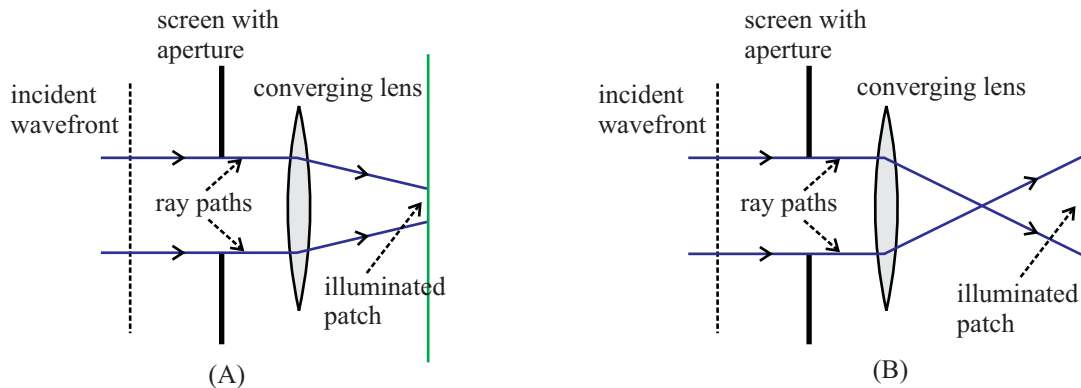


Figure 15-17: Modification of the set-up of fig. 15-3 where the observation screen is (A) in front of and (B) behind the focal plane of the converging lens; in either case, the rules of ray optics predict the formation of an illuminated patch on the screen surrounded by a shadow region; however, the bending and spreading of the wave (referred to as diffraction in the present context) results in the formation of a fringe pattern similar to the one in fig. 15-2(B), described as a Fresnel pattern; the special case of the observation screen being placed in the focal plane of the lens corresponds to the Fresnel pattern being reduced to a Fraunhofer one.

15.4.5 The single slit Fraunhofer pattern

Figure 15-18 is a set-up for Fraunhofer diffraction by a narrow slit (in the screen S_1) of width a , where a plane wave is made to be incident on the slit from the left, and the diffraction pattern is observed on the screen S_2 , where S_2 is to be imagined to be at an *infinitely large distance* from the slit. Note that there is no lens in this figure to the right of S_1 as in fig. 15-3. Fig. 15-19(A) shows a section of an *equivalent* arrangement by a plane perpendicular to the length of the slit, where now a converging lens is placed to the right of the slit, and the observation screen S_2 is placed *in the focal plane of the lens*.

15.4.5.1 Ray paths corresponding to secondary waves.

Ray paths from the left representing wave normals of the incident plane wave give rise to ray paths (corresponding to secondary waves) originating from all the points in the aperture, where these ray paths correspond to the secondary waves in the diffraction region, i.e., to the right of S_1 . Referring first to fig. 15-18, a point P on S_2 can receive only those secondary waves from the slit whose ray paths are parallel so that they can ‘meet at infinity’. What the lens does is to collect all these parallel rays so that these meet at a point in the focal plane. Two such sets of parallel ray paths are shown in

fig. 15-19(A), of which one set corresponds to undeviated rays, focused at the point O, while the other are deviated by an angle θ to the direction of the incident rays, focused at the point P.

Note that the incident rays in this set-up may be assumed to come from a point object located at an infinitely large distance from the slit on the axial line $O'O$, and the point O in fig. 15-19(A) is the *image*, according to the rules of geometrical optics, of this distant point object. The other sets of deviated ray paths, like the one shown to converge to P due to the action of the lens, all arise because of the fact that the incident wave gets fanned out, on both sides of the forward direction, in moving past the slit.

Since the action of the lens is to collect and focus ray paths that would otherwise meet at infinity, one can describe the Fraunhofer pattern as a special instance of a Fresnel pattern, namely, one that is formed at an infinite distance from the aperture (in situations where the source is located effectively at an infinitely large distance and, moreover, no image forming device such as a lens is used). However, a more general description of a Fraunhofer pattern is to say it is produced in a conjugate plane to the source of light, i.e., in the plane where the geometrical image of the source is formed.

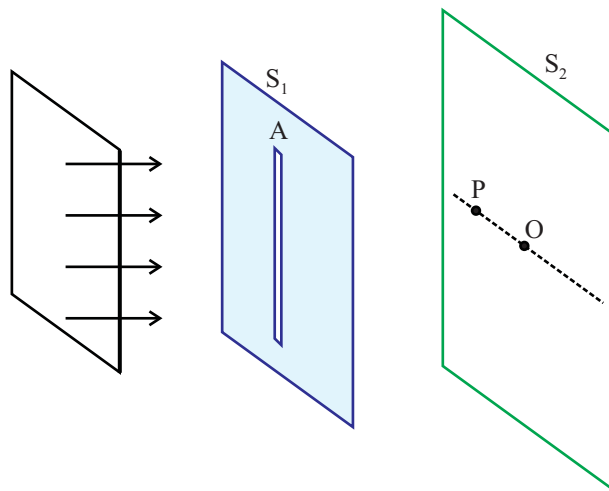


Figure 15-18: Fraunhofer set-up with a narrow slit, with no lens to the right of the slit; instead, the observation screen is imagined to be placed at an infinite distance to the right of the slit.

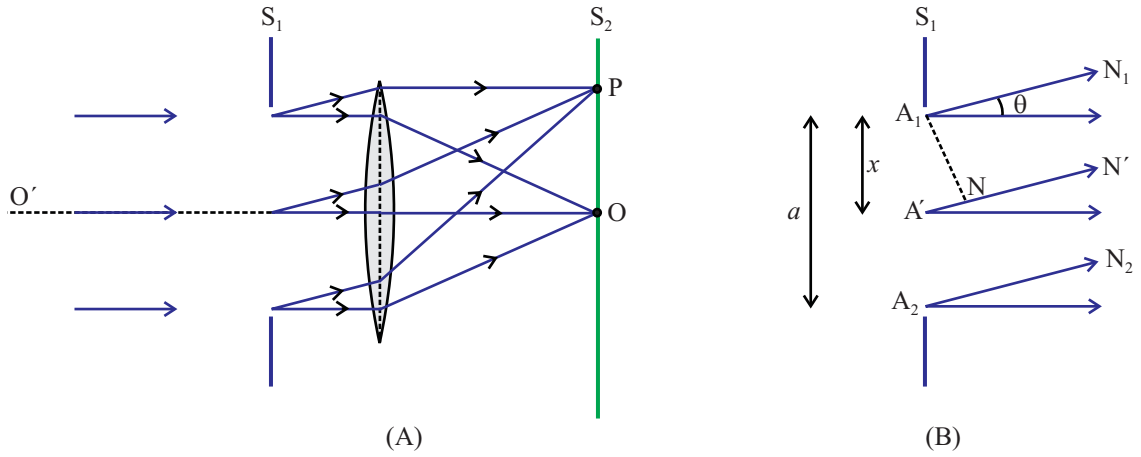


Figure 15-19: (A) Section by a plane perpendicular to the length of the slit shown in fig. 15-18, now with a lens to the right of the slit and the observation screen placed at the focal plane of the lens; O is the geometrical image of the infinitely distant source point, while P is any other observation point on the screen; (B) two parallel bunches of diffracted ray paths, one to O and the other to P ; among the latter, A_1N_1 and $A'N'$ are two ray paths, one from the point A_1 located at one end of the slit, and other from from A' at a distance x from A_1 ; A_1N is dropped perpendicular on $A'N'$.

15.4.5.2 The intensity formula.

The intensity at P can be calculated by making use of the basic formula (15-34), for the purpose of which refer to fig. 15-19(B). A set of parallel ray paths deviated from the forward direction by an angle θ is shown here. The ray path A_1N_1 originates from one end (A_1) of the slit while $A'N'$ originates from the point A' at a distance x from A_1 . Considering an element of width δx around the point A' , the contribution of this element to the complex wave function is proportional to $e^{ikl}\delta x$ where l is the optical path length from A' to P (the point of observation on the screen S_2).

Let l_0 be the optical path length from A_1 to P . Then, looking at fig. 15-19(B), one can write $l = l_0 + x \sin \theta$. Thus, summing up contributions to the wave function at P from all such elements from A_1 to A_2 , and then taking the modulus squared of the resulting integral, one obtains the intensity I at P :

$$I = N \left| e^{ikl_0} \int_0^a e^{ikx \sin \theta} dx \right|^2, \quad (15-35)$$

where N is a multiplicative constant which is required to obtain the intensity from the expression in eq. (15-34), the latter being the intensity on a relative scale.

Evaluating the integral, one ends up with the expression

$$I = I_0 \left(\frac{\sin \beta}{\beta} \right)^2, \quad (15-36a)$$

where

$$\beta = \frac{ka}{2} \sin \theta = \frac{\pi a}{\lambda} \sin \theta. \quad (15-36b)$$

Note that, with $\theta = 0$ in eq. (15-36a), (15-36b), one gets $I = I_0$, i.e., in other words, the constant I_0 stands for the intensity at the point O (the geometrical image point, corresponding to $\theta = 0$, the forward direction in the present context).

Problem 15-8

Establish eq. (15-36a) from (15-35).

Answer to Problem 15-8

HINT: $\int_0^a e^{ikx \sin \theta} dx = \frac{1}{ik \sin \theta} (e^{ika \sin \theta} - 1) = \frac{e^{i \frac{ka}{2} \sin \theta}}{ika \sin \theta} (e^{i \frac{ka}{2} \sin \theta} - e^{-i \frac{ka}{2} \sin \theta}) = e^{i \frac{ka}{2} \sin \theta} \frac{\sin(\frac{ka}{2} \sin \theta)}{\frac{ka}{2} \sin \theta}$. Multiplying with e^{ikl_0} and taking the modulus squared we obtain $I = N \left(\frac{\sin \beta}{\beta} \right)^2$, where β is given by (15-36b). Eq. (15-36a) then results by a renaming of the constant N .

15.4.5.3 Absence of diffraction in the vertical direction.

An important observation to make here is that, if the slit A is sufficiently long, then *there occurs no diffraction in a direction parallel to the length of the slit*, i.e., the diffracted ray paths fan out only in horizontal planes perpendicular to the length of the slit. Fig. 15-20 shows the ray paths corresponding to the secondary waves in three vertical planes, of which the plane V_2 contains the central ray path O'O (see fig. 15-19(A)), where one finds

that the ray paths fan out in the horizontal direction on either side of $O'O$, but not in the vertical direction.

The reason why all the ray paths in any of the vertical planes (V_1, V_2, V_3 in the figure) are parallel, as in the incident bunch of rays, is that the length of the slit being very large compared to the wavelength λ of light (typically, of the order of 10^6 times), there does not occur any bending and spreading of the wave in a vertical plane, and ray paths in this plane follow the rule of rectilinear propagation of geometrical optics. On the other hand, the dimension of the slit (a) in the horizontal direction being much smaller (of the order of 10^3 times the wavelength in a typical diffraction set-up), there occurs a fanning out of the ray paths on either side of the central ray path OO' in the horizontal plane.

Since there is no diffraction in a vertical plane, all rays-paths in the diffraction region are collected by the lens (placed in the plane L in fig. 15-20) along the line Q_1Q_2 in the focal plane of the lens. The expression (15-36a) then gives the intensity distribution along this line as a function of the angle of diffraction θ , where positive and negative values of θ correspond to rays deviated on the two sides of the central ray path OO' (we have taken θ to be positive for ray paths deviated to the left of the forward direction, as in fig. 15-19(B)).

In this context, one notes that the expression (15-34) involves a *surface* integration over the aperture, while eq (15-35) involves an integration only over the *width* of the aperture. Strictly speaking, one needs to integrate over the length of the aperture as well, but when this length is sufficiently large (ideally, *infinitely* large) the result of that integration gives a non-zero value only for points on the line $O'O$. The remaining integration over the width of the aperture (expression (15-35)) leads to the intensity distribution formula (15-36a).

15.4.5.4 The intensity graph

Fig. 15-21 gives a graphical plot of the variation of the intensity I with angle of diffraction θ , in accordance with formula (15-36a). One observes that the intensity distribution

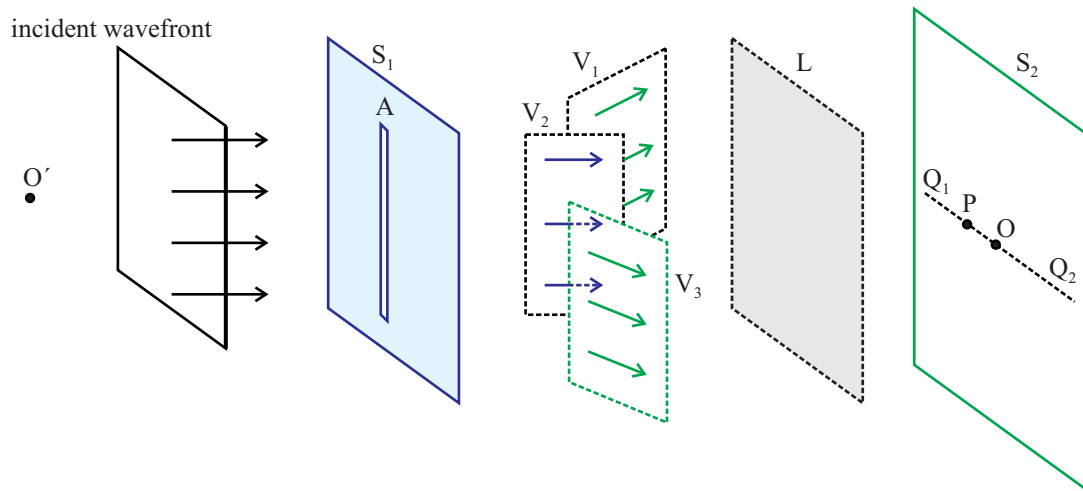


Figure 15-20: Showing the fanning out of ray paths in a horizontal plane but no diffraction in a vertical plane; three vertical planes (V_1 , V_2 , V_3) containing diffracted ray paths corresponding to secondary waves originating in the aperture are shown; due to the length of the diffracting slit being large compared to the wavelength of light used, all diffracted ray paths are collected by the lens, placed in the plane L , on to the line Q_1Q_2 in its focal plane; one thereby gets a variation of intensity along the line Q_1Q_2 in accordance with formula (15-36a).

consists of *maxima* and *minima*, with the central maximum at $\theta = 0$, followed by other maxima of successively smaller values of intensity, along with intervening minima of zero intensity. These maxima and minima may be interpreted as being produced by *constructive* and *destructive* interference among the secondary waves that may be imagined to have originated at various different points in the slit-like aperture.

The maxima and minima may be assigned *order numbers*, with the central maximum being one of order $n = 0$, while other maxima on either side of this central maximum correspond to $n = \pm 1$, $n = \pm 2$, Incidentally, the central maximum corresponds to the *geometrical image* of the point source, located at an infinitely large distance, produced by the converging lens. The effect of the slit is then to produce a fringing effect of this geometrical image by the maxima and minima on either side.

The minima may also be similarly ordered on either side of the central maxima as $n = \pm 1$ (first order), $n = \pm 2$ (second order), and so on. Formula (15-36a) tells us that the n th

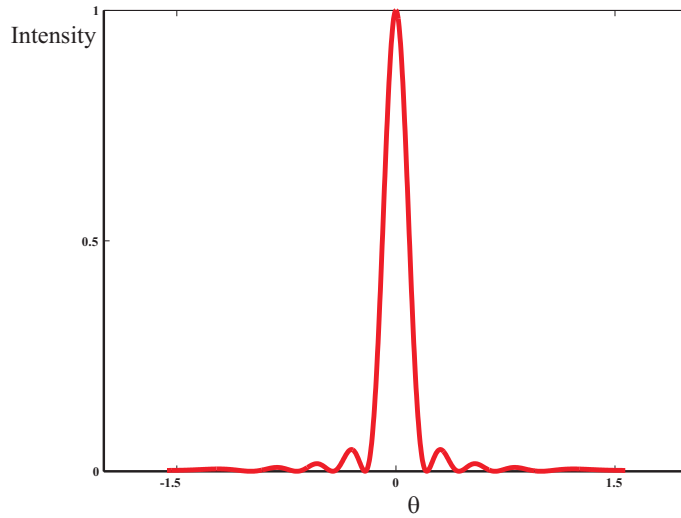


Figure 15-21: Variation of intensity with angle of diffraction (θ) in a single slit diffraction pattern; the intensity distribution consists of a central maximum at $\theta = 0$, on either side of which there are a number of other maxima of smaller intensity, separated by minima of zero intensity; for the purpose of presentation, the intensity has been scaled by I_0 , the intensity of the central maximum, so that the latter now becomes unity; the entire distribution now depends on the single parameter $\frac{\pi a}{\lambda}$ (see (15-36b)), which has been chosen to be 15.

order minimum corresponds to a value of θ given by

$$\beta = n\pi, \text{ i.e., } a \sin \theta = n\lambda. \quad (n = \pm 1, \pm 2, \dots) \quad (15-37)$$

Fig. 15-22 depicts the effect of the *slit width* on the intensity distribution, where the distributions are compared for two slits with different values of the parameter $\frac{\pi a}{\lambda}$. As the width of the slit is made to increase, the fringes get narrower since more and more light gets concentrated on to the geometrical image (the central bright ‘fringe’, i.e., the one of order $n = 0$), following the rules of geometrical optics, and the effect of bending and spreading of the wave becomes comparatively less pronounced. On the other hand, with a *narrower* slit, the wave gets bent and spread away from the forward direction to a greater degree, and the fringes spread out.

Problem 15-9

In the Fraunhofer pattern produced by a single slit of width $a = 0.03 \text{ mm}$, the intensity at the central maximum is $I_0 = 1.5 \times 10^{-3} \text{ W}\cdot\text{m}^{-2}$, the slit being illuminated normally with a plane wave

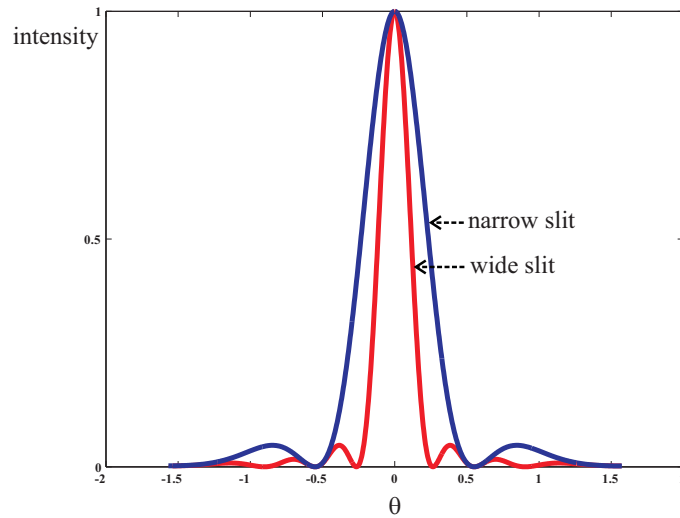


Figure 15-22: Illustrating the effect of slit width on the intensity distribution in the single slit Fraunhofer diffraction pattern; for a given wavelength, a wider slit corresponds to comparatively narrow fringes, where most of the light is concentrated in the central fringe formed around the geometrical image; for a narrower slit with a smaller value of the parameter $\frac{\pi a}{\lambda}$, the central fringes gets spread out, corresponding to a greater degree of bending and spreading of the wave as it passes the slit; the values of $\frac{\pi a}{\lambda}$ chosen for the two graphs are 12 (inner), and 6 (outer).

of wavelength $\lambda = 500 \text{ nm}$. What is the intensity at a diffraction angle $\theta = 0.02^\circ$? What will the intensity at the same diffraction angle be if the slit width is changed to $a = 0.035 \text{ mm}$?

Answer to Problem 15-9

Making use of the formula (15-36a) one finds, for $a = 0.03 \text{ mm}$, $\beta = \frac{\pi a}{\lambda} \sin \theta \approx 1.2\pi$, i.e., $I \approx 0.0244I_0 = 3.66 \times 10^{-5} \text{ W}\cdot\text{m}^{-2}$. On the other hand, with $a = 0.035 \text{ mm}$, β changes to 1.4π (approx), and thus $I = 0.0466I_0 = 6.99 \times 10^{-5} \text{ W}\cdot\text{m}^{-2}$ (approx).

15.4.5.5 Fraunhofer fringes with a slit-source.

Interference and diffraction patterns are seldom observed with a point source since the fringes become indistinct due to scanty light from the source. In the Fraunhofer set-up with a long narrow diffracting slit and a point source of light, variations of intensity along a single line do not produce a distinct visual impression. In the above paragraphs I have referred to a single plane wave front incident on the aperture or a slit only for the sake of convenience.

A point source placed at the first focal point of a converging lens produces a plane wave. Such a lens is referred to as a collimating lens. An alternative way to look at a plane wave is to imagine that the bunch of parallel rays corresponding to such a wave originates from an infinitely distant point source.

Fig. 15-23 shows a Fraunhofer set-up with a vertical *slit-source* S placed in the focal plane of a converging (the *collimating*) lens C, while the rest of the set-up looks the same as in fig. 15-3 (with a long narrow slit replacing the circular diffracting aperture). The use of the slit source leads to the formation of Fraunhofer fringes in the form of alternate dark and bright lines on either side of the central bright fringe which, in reality, is the image of the slit source formed by the combination of the collimating lens C and the lens L placed to the right of the diffracting slit. Each fringe is characterized by an angle of diffraction θ , and a corresponding value of $\beta = \frac{\pi}{\lambda} a \sin \theta$, where the values of β for the dark fringes of various orders are given by (15-37).

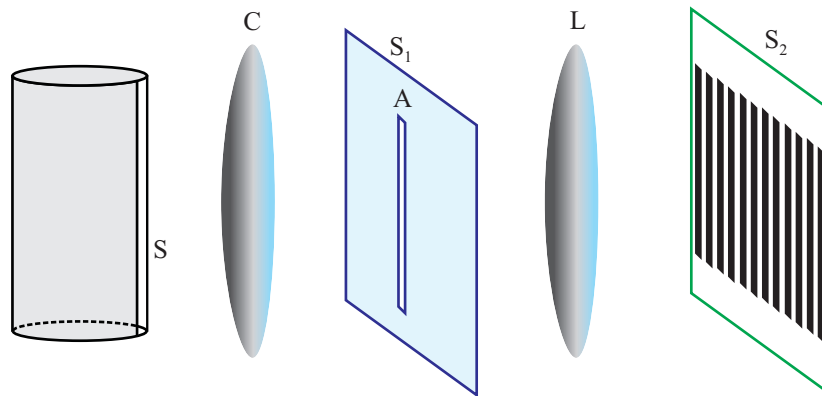


Figure 15-23: Single slit Fraunhofer set-up with a slit source S and a collimating lens C; straight line fringes on observation screen S_2 are formed on either side of the central fringe (not marked in the figure), the latter being the geometrical image of the slit source.

15.4.5.6 Phase in Fraunhofer diffraction

Recall that the complex representation of a wave function for a harmonically varying field at any given point in space is often of the form $Ae^{i\Phi}e^{-i\omega t}$ where A does not depend strongly on the position and may include a constant phase while Φ is the space-

dependent part of the phase and varies rapidly with the location of the point under consideration. The product $Ae^{i\Phi}$ is commonly referred to as the (complex) amplitude of the wave function at the given point with reference to the time-dependent phase factor $e^{-i\omega t}$.

Formula (15-34) then involves, first of all, a summation of the complex amplitudes corresponding to secondary waves reaching a given field point, and then, taking the modulus squared to arrive at the intensity.

What is crucial importance in this context is that, the typical space-dependent phase $\Phi = kx \sin \theta$ is *linear* in the co-ordinate x specifying the location in the aperture (the diffracting slit in the present context) of the point from which a secondary wave originates. In the case of Fresnel diffraction, on the other hand, the expression for the phase involves terms of the second and higher degrees in the co-ordinates of the point of origin of the secondary wave.

This makes the task of working out the sum of amplitudes much easier for Fraunhofer diffraction compared to calculating the corresponding amplitude sum in the case of Fresnel diffraction. This simplifying feature of Fraunhofer diffraction is related to the fact that the Fraunhofer pattern is assembled in the plane of the geometrical image of the source, formed by intervening lens systems.

15.4.5.7 Coherence properties and diffraction fringes

The idea of coherence of an electromagnetic wave was introduced in section 14.10. Coherence is a broad concept where one may consider the coherence characteristics of a single electromagnetic signal as also of the *mutual* coherence properties of two different signals.

As indicated in section 14.10, there are several aspects to the coherence characteristics of an electromagnetic wave such as, for instance, its degree of monochromaticity and its *polarization* characteristics (polarization properties of electromagnetic and optical waves are discussed in sections 14.4.8 and 15.5).

Let us first consider the formation of diffraction fringes with *polarized light* from a point source, in which case a lack of coherence arises because of deviation from strict monochromaticity, and the degree of coherence can be quantitatively indicated in terms of the *coherence length* of the signal under consideration. In this case the criterion for the formation of diffraction fringes is that the path difference between the secondary waves reaching a field point has to be less than the coherence length of the light (more specifically, one has to consider here the values of path difference corresponding to waves from various different pairs of points on the aperture).

Imagining, on the other hand, a perfectly monochromatic point source emitting unpolarized or partially polarized light, the lack of coherence on *this* count does not adversely affect the formation of diffraction fringes. One can decompose the unpolarized light signal into two plane polarized components, with phase difference between the two constituting a random variable. Each of the two components, considered by itself, then results in an intensity distribution arrived at by the *sum of amplitudes* formula. One then employs the *sum of intensities* formula in summing up the two intensities at each field point due to the two components so as to arrive at the resultant intensity distribution. Since the two polarized components lead to identical intensity distributions, their sum also produces an intensity distribution of the form (15-36a).

In other words, the effect of the coherence characteristics of light on the formation of diffraction fringes is similar to that in the case of interference fringes (refer back to sections 15.3.6.1 and 15.3.6.2).

15.4.6 The double slit Fraunhofer pattern

Fig. 15-24 depicts a Fraunhofer set-up with a slit source and a collimating lens, where the diffraction is caused by *a pair of* parallel narrow slits in the screen S_1 , followed by the converging lens L and the observation screen S_2 in the focal plane of the lens. The figure shown is a section of the set-up by a plane perpendicular to the length of the diffracting slits.

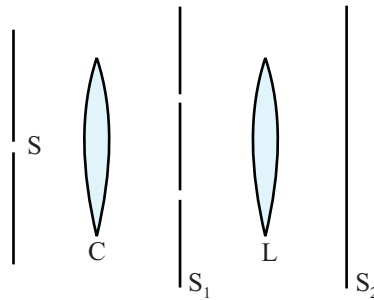


Figure 15-24: Double slit Fraunhofer set-up with a slit source; a section by a plane perpendicular to the length of the slit source and to the diffracting slits (which are assumed to be all parallel) is shown; the slit S is illuminated with monochromatic light, and it acts as a collection of independent point sources.

Fig. 15-25 shows a bunch of parallel ray paths, corresponding to secondary waves originating from points on the two slits, where the bunch is made up of two groups of ray paths, one from each slit, all the ray paths shown being deviated from the forward direction by a diffraction angle θ (the ray paths shown here are all deviated to the right of the forward direction, in contrast to those in fig. 15-19; the intensity distribution patterns are, however, symmetric about the forward direction). All these rays are collected by the converging lens on the observation screen S_2 , and a set of alternating dark and bright straight line fringes is formed in this plane, as in the case of the single slit pattern with a slit source.

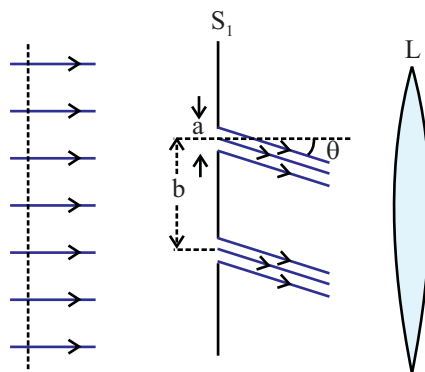


Figure 15-25: Formation of the double slit Fraunhofer pattern; showing a bunch of parallel ray paths corresponding to secondary waves; the bunch is made up of two groups, one from each slit; the ray paths are all contained in a plane perpendicular to the lengths of the slits; however, for a slit source, ray paths lying in planes making various angles with this plane are also relevant in the formation of straight line fringes.

Considering all ray paths with a given diffraction angle θ , one can once again make use of the general formula (15-34) to arrive at an expression for intensity similar to (15-35), but now the range of integration breaks up into two parts corresponding to the two slits:

$$I = N \left| \int_0^a e^{ikx \sin \theta} dx + \int_b^{a+b} e^{ikx \sin \theta} dx \right|^2, \quad (15-38a)$$

Here a stands for the width of either slit and b for the separation between the slits, both the slits being assumed identical for the sake of simplicity. On working out the integrals, this expression is seen to reduce to the form

$$I = I_0 \cos^2 \frac{\delta \sin^2 \beta}{2 \beta^2}, \quad (15-38b)$$

where β is given by eq. (15-36b), and δ by

$$\delta = \frac{2\pi}{\lambda} b \sin \theta, \quad (15-38c)$$

while the constant I_0 stands for the intensity corresponding to the angle $\theta = 0$ (the ‘central fringe’, resulting from the undeviated rays). The intensity distribution of the double slit Fraunhofer pattern, given by the above expression, is shown schematically in fig. 15-26.

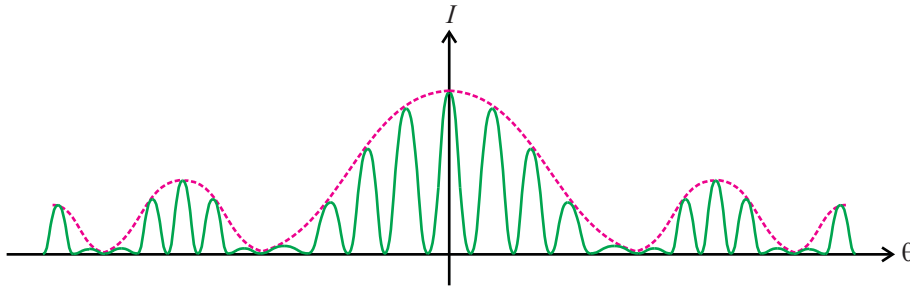


Figure 15-26: I - θ graph (schematic) for the double slit Fraunhofer pattern, given by formula (15-38b); while the solid curve depicts the intensity distribution, the dotted curve represents its envelope, and represents the single slit diffraction pattern, analogous to the one in fig. 15-21.

Equations (15-38a) and (15-38b) admit of the following interesting interpretation. One

can write eq. (15-38a) in the form $C |\mathcal{A} + e^{i\delta} \mathcal{A}|^2$, where \mathcal{A} stands for the integral in the expression (15-35). This can be interpreted as the intensity (up to a multiplicative constant) resulting from the *interference* of two waves, each with a complex amplitude \mathcal{A} , and with a phase difference δ between the two.

Here \mathcal{A} represents the complex amplitude of the signal sent out to the field point by each of the two slits considered in isolation, while δ stands for the phase difference between the two signals. Indeed the factor $\cos^2 \frac{\delta}{2}$ in eq. (15-38b) resembles the analogous factor in the intensity expression (15-3) that describes a double slit interference pattern, while the factor $\frac{\sin^2 \beta}{\beta^2}$ gives the intensity distribution (again up to a factor) in the single slit diffraction pattern.

In other words, the double slit diffraction phenomenon can be looked upon as the interference of two optical signals, each produced by a single slit diffraction. This is borne out by the intensity distribution graph of fig. 15-26 which corresponds to the intensity graph of a double slit interference pattern (fig. 15-8), but with the intensity of the maxima *modulated* by the single slit intensity graph of the type of fig. 15-21, the latter corresponding to a relatively slow variation of intensity with angle θ . The single slit diffraction graph (dotted line in fig. 15-26) acts here as the *envelope* of the double slit interference graph.

The angle θ corresponding to the first diffraction minimum on either side of the central maximum is given by

$$a \sin \theta = \pm \lambda, \quad (15-39a)$$

while the first interference minimum on either side of the central maximum is given by

$$b \sin \theta = \pm \frac{\lambda}{2}. \quad (15-39b)$$

The interference maxima can be labelled with the order number n , where the n th maxi-

mum is given by

$$b \sin \theta = n\lambda \quad (n = \pm 1, \pm 2, \dots). \quad (15-39c)$$

The double slit fringes take the appearance of successive diffraction *bands*, where each band is made up of a number of narrower interference fringes.

Problem 15-10

Establish eq. (15-38b).

Answer to Problem 15-10

HINT: In the expression $\int_0^a e^{ikx \sin \theta} + \int_b^{a+b} e^{ikx \sin \theta}$, one can replace the integration variable x in the second integral with a new variable $b + x'$ and then rename x' as x , which gives the expression $\mathcal{A} + e^{ikb \sin \theta} \mathcal{A}$, where $\mathcal{A} = \int_0^a e^{ikx \sin \theta} dx$. This is precisely of the form of the integral occurring in formula (15-35), which was evaluated in problem (15-8) to $e^{i\beta \frac{\sin \theta}{\beta}} (\beta = \frac{ka}{2} \sin \theta = \frac{\pi}{\lambda} a \sin \theta)$. Defining δ as in (15-38c), we obtain, from (15-38a), $I = N |e^{i\beta \frac{\sin \theta}{\beta}} (1 + e^{i\delta})|^2$. Now, $|1 + e^{i\delta}|^2 = 2 + 2 \cos \delta = 4 \cos^2 \frac{\delta}{2}$. This gives eq. (15-38b), with $I_0 = 4N$, which can be interpreted as the intensity for $\theta = 0$.

Problem 15-11

How many bright double-slit interference fringes are formed between the first diffraction minima on the two sides of the central maximum in a double slit Fraunhofer diffraction pattern formed by monochromatic light with wavelength $\lambda = 600$ nm, if the width of either slit is $a = 0.04$ mm and the separation between the slits is $b = 0.2$ mm? What is the ratio of the intensity of the first bright fringe on either side of the central fringe, to the intensity of the latter? .

Answer to Problem 15-11

The intensity distribution in a single slit diffraction pattern (with slit width a) is given by the factor $\frac{\sin^2 \beta}{\beta^2}$ in formula (15-38b), which corresponds to the envelope in fig 15-26. The diffraction angle corresponding to the first diffraction minimum on either side of the central maximum is given by $a \sin \theta = \pm \lambda$, i.e., $\theta = \pm \sin^{-1}(\frac{\lambda}{a}) = \pm 0.015$ (approx). Within the range between these two

angles, the interference maxima, given by the factor $\cos^2(\frac{\delta}{2})$ in eq. (15-38b), are obtained from the formula (15-39c).

Thus, with $\theta = 0.015$, one gets $n = \frac{b \sin \theta}{\lambda} = 5$. In other words, the number of interference maxima on either side of the central maximum is 5. However, the fifth maximum on either side is suppressed since it coincides with the first diffraction minimum (the product of the two factors in eq. (15-38b) gives $I = 0$; such suppression of an interference maximum is said to correspond to a ‘missing maximum’). In other words, the required number of intensity maxima, counting the central maximum, is 9.

The first interference maximum on either side of the central maximum corresponds to $b \sin \theta = \pm \lambda$, i.e., $\theta = \pm 0.003$. For each of these, the factor $\cos(\frac{\delta}{2})$ is unity, i.e., the same as that for the central maximum. Hence the required ratio is given by $\frac{\sin^2 \beta}{\beta^2}$ corresponding to the above value of θ (i.e., to $\beta = \frac{\pi}{\lambda} a \sin \theta = \frac{\pi \times 4 \times 10^{-5} \times 3 \times 10^{-3}}{6 \times 10^{-7}} = \frac{2\pi}{10}$; refer back to eq. (15-38b); the value of the factor $\frac{\sin^2 \beta}{\beta^2}$ for the central maximum, i.e., for $\beta \rightarrow 0$, is unity). In other words, the required intensity ratio is $\frac{\sin^2(\frac{2\pi}{10})}{(\frac{2\pi}{10})^2} = 0.875$ (approx).

15.4.7 The diffraction grating

The diffraction grating is an extension of the double slit arrangement where one has a *large number* of parallel narrow slits cut side by side in a screen, with opaque spaces in between successive slits. The slits and the opaque spaces between them are made very narrow indeed, so that gratings with something like five thousand lines (i.e., transparent slits) per cm are quite common. The slits are formed by special etching techniques, usually on a transparent surface with an opaque coating, and are referred to as ‘rulings’ on the grating surface. Gratings with their etchings imprinted on curved surfaces are also possible.

The Fraunhofer pattern formed by a diffraction grating with a slit source and a collimating lens consists a number of sharp bright fringes with appreciably large gaps separating them, the gaps being made of dark spaces where the intensity is almost zero. The intensity distribution graph in the Fraunhofer pattern looks as in figure 15-27 where one

finds, in between the widely separated sharp maxima, a number of secondary maxima where the intensities are small in magnitude, being almost negligible compared to those of the principal maxima.

Carrying forward the argument from where we wrote down the formula (15-38a), the expression for the intensity distribution in a grating Fraunhofer pattern can be written in the form

$$I = C \left| \mathcal{A} + \mathcal{A}e^{i\delta} + \mathcal{A}e^{2i\delta} + \cdots + \mathcal{A}e^{i(N-1)\delta} \right|^2, \quad (15-40a)$$

(with a slight change in notation, where the constant N has been replaced with C , the symbol N now standing for the total number of slits) where

$$\mathcal{A} = \int_0^a e^{ikx \sin \theta}, \quad (15-40b)$$

and

$$\delta = \frac{2\pi}{\lambda} b \sin \theta. \quad (15-40c)$$

In these equations N stands for the total number of transparent slits constituting the grating, a for the width of each slit, b for the distance between the mid-points of successive slits, and θ for the angle of diffraction. Notice that, in the above formula, there occurs a sum corresponding to the interference of the signals sent out to the field point by all the N slits, the phase difference between the signals from successive slits being δ .

On working out the expressions (15-40a), (15-40b), one arrives at the following formula for the intensity distribution

$$I = \frac{I_0}{N^2} \frac{\sin^2 \frac{N\delta}{2}}{\sin^2 \frac{\delta}{2}} \frac{\sin^2 \beta}{\beta^2}, \quad (15-41a)$$

where

$$\beta = \frac{\pi a}{\lambda} \sin \theta, \quad (15-41b)$$

and from which follows figure 15-27. Illustrating what I have already mentioned above, this graph shows principal maxima of a few orders ($n = 0, \pm 1, \pm 2$), with secondary maxima in between. In the above expression, I_0 once again stands for the intensity of the central maximum ($n = 0$), i.e., in the forward direction, $\theta = 0$. The dotted curve is the envelope corresponding to the single slit factor $\frac{\sin^2 \beta}{\beta^2}$, showing only a part of the central diffraction maximum, where the intensity falls off very slowly because of the fact that the slits in the grating are made very narrow.

The diffraction angle for the principal maximum of order n ($n = 0, \pm 1, \pm 2, \dots$), is given by (refer to eq. (15-41a))

$$b \sin \theta = n\lambda. \quad (15-42)$$

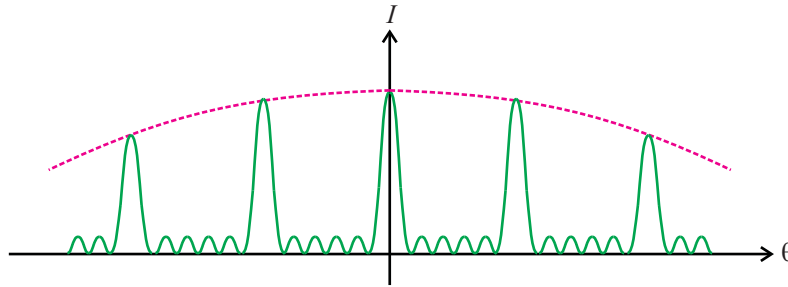


Figure 15-27: I - θ graph for diffraction grating showing principal maxima of a few orders and secondary maxima (schematic); dotted curve shows envelope corresponding to single slit diffraction; the intensity distribution, given by formula (15-41a) consists of sharp maxima separated by numerous feeble maxima, with minima in between.

Thus, the positions of the principal maxima of various orders depend on the wavelength of the light used. This, along with the fact that the principal maxima are sharp and widely separated, makes the grating a useful device for effecting spectral separation of the frequency components in an optical signal made up of more than one such components, and for the accurate determination of the wavelengths.

What I have referred to above is the intensity distribution formed by the grating in the *transmitted* optical field. Analogous intensity patterns in the *reflected* field are

also possible. Moreover, it has been assumed that the incident plane wave strikes the grating surface normally, i.e., the incident ray paths are normal to the grating surface. The analysis can be extended without much difficulty to the case when the incident ray paths are inclined to the grating surface at some other angle.

Problem 15-12

A grating having 300 rulings per mm is used to form diffraction fringes with monochromatic light of wavelength λ incident normally on it. If diffraction fringes are formed with the transmitted light, calculate the maximum wavelength for which the sixth order principal maximum can be formed with the grating. What is the maximum order of the diffraction maximum that can be formed with visible light, of wavelength ranging from 300 to 800nm?

Answer to Problem 15-12

The principal maxima in the diffraction pattern formed by the grating are given by the formula (15-42) in which we have to take $b = \frac{1}{300 \times 10^3}$ m (reason this out) and $n = 6$ ($n = -6$ corresponds to a negative value of θ , symmetrically related to $n = 6$). The required maximum wavelength λ_0 (say) is obtained by taking $\theta = \frac{\pi}{2}$, the maximum possible diffraction angle for transmitted light. Thus, $\lambda_0 = \frac{b}{6} = 555.5$ nm (approx). Conversely, with $\theta = \frac{\pi}{2}$, the maximum order of principal maximum corresponds to the minimum wavelength in the visible range. However, with $\lambda = 300$ nm, one gets $N = \frac{1}{300 \times 10^3 \times 300 \times 10^{-9}} = 11.11$, which is a fractional number. A larger integral value of N would correspond to a wavelength below the minimum visible wavelength. Thus, the maximum order of principal maximum is $N = 11$, corresponding to $\lambda = \frac{b}{11} = \frac{1}{300 \times 10^3 \times 11}$ m = 303 nm (approx).

15.4.8 Resolving powers of optical instruments

Optical instruments like the telescope and the microscope were introduced in section 10.16. While the telescope forms the image of a distant object by bringing about *angular magnification*, the microscope, on the other hand, introduces *linear magnification* in the imaging of a small object placed close to its objective lens. In both of these

imaging systems, special care is taken so as to eliminate aberrations of various kinds as completely as possible in order that the images replicate the respective objects to a sufficiently high degree. While residual aberrations continue to affect the quality of imaging, a more serious problem is often posed by the *diffraction* caused by the lenses used in the imaging systems.

Fig. 15-28 depicts schematically how a lens, while imaging a distant point object (say, O), itself causes a bending and spreading of the wave incident on it, where this bending and spreading occurs predominantly at the periphery of the lens. This is a diffraction effect caused by the lens itself acting as the diffracting object which, in the present context, is analogous to a circular aperture. As a result of this diffraction, there occurs a loss of definition of the image (I) produced by the lens since a *Fraunhofer pattern* is formed around the image, causing the latter to lose its sharpness. Thus, if a screen is placed at the focal point of the lens as shown in the figure, then alternate dark and bright circular fringes are produced around the image point, thereby causing the image to become blurred.

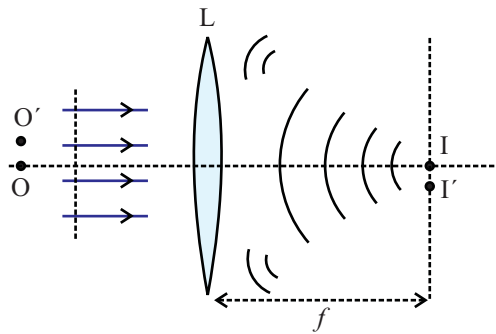


Figure 15-28: Diffraction caused by a lens due to finiteness of lens aperture; spreading and bending of incident wave, due to which the image I of a distant point object O, formed at the focal point, is surrounded by alternate dark and bright circular fringes; O' is a second point object at a small angular separation from O whose image I' will be similarly fringed, see fig. 15-29(A),(B).

Let us now imagine, instead of one single point object, a *pair* of distant point objects at a small angular separation as seen from the center of the lens, where the second object (O' in the figure) is similarly imaged by the lens at the point I', say, (see fig. 15-

29(A), (B)) close to I, and where both I and I' are fringed with their respective Fraunhofer patterns. If the angular separation between O and O' (fig. 15-28) be sufficiently large, then the images I and I' are also well separated from each other, and the formation of the Fraunhofer patterns by the lens does not stand in the way of the observer identifying the images distinctly. If, on the other hand, the angular separation be small, then I and I' come so close together that the Fraunhofer patterns around them make them appear as one single blurred patch, and the objects O and O' then fail to be *resolved* by the imaging system (a single lens in the present instance).

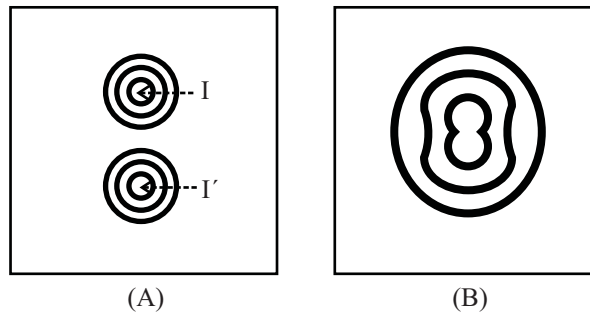


Figure 15-29: Images I and I' of two point objects (O and O' in fig. 15-28) formed by a lens; (A) the images are well separated and can be distinguished from each other, in spite of each being blurred due to the fringing effect; (B) the images are so close together that they fail to be distinguished from each other.

This is a simple example of how the diffraction caused by lenses in imaging systems used in optical instruments may stand in the way of resolving the structural details of a composite object, such as a binary star observed through a telescope or a bacterial cell observed by a microscope. The angular or linear dimension of the smallest structure that can be resolved by an instrument is referred to as its *resolving power*.

The idea of resolving power is also relevant in the context of spectrally separating the components, with frequencies close to one another, of quasi-monochromatic light (i.e., radiation contained within a small range of frequencies) emitted from a source. For instance, the bright lines corresponding to the principal maxima of a certain order, formed by a diffraction grating, for two frequencies, say, ν and $\nu + \delta\nu$, may be formed so close to each other that their separation becomes comparable to the width of each of

the lines, when they may fail to be distinguished as two distinct lines. This then means that the grating fails to *resolve* these two frequencies.

15.5 Polarized and unpolarized light

Section 14.4.8 dealt with the concept of the state of polarization of a plane monochromatic electromagnetic wave. One needs to specify the state of polarization, in addition to the angular frequency and the direction of propagation, so as to completely describe a plane monochromatic wave. The fundamental distinction in this context is between *polarized* and *unpolarized* waves, which can be explained by looking at the wave under consideration as a superposition of a pair of component waves, where each is linearly polarized in a direction perpendicular to the other, as explained in sec 14.4.8. The crucial feature of relevance here is the *amplitudes* of and the *phase difference* between the two component waves. Let us, then, briefly recall the basic ideas following which the various states of polarization can be related to these characteristics.

15.5.1 The basic components: x-polarized and y-polarized light

The two basic linearly polarized states are depicted in fig. 15-30(A) and (B) where, in (A) the amplitude of the electric field intensity is directed along the x-axis while in (B), the amplitude is along the y-axis of a right handed Cartesian co-ordinate system, and we assume that the wave propagates along the positive z-axis. We refer to these as the *x-polarized* and the *y-polarized* components respectively.

With the above specification of the direction of propagation and of the electric field vectors one need not specify separately the respective directions of the magnetic field intensity since, as we have already seen (refer to section 14.4.1), the electric field intensity, the magnetic field intensity, and the unit vector along the direction of propagation form a right-handed orthogonal triad.

15.5.2 Specifying the basic components and their phase relation

For the sake of reference, the electric field vectors for the basic x-polarized and y-polarized waves, with no phase difference between them, will be assumed to be of the form $\hat{i}Ae^{-i\omega t}$ and $\hat{j}Be^{-i\omega t}$ respectively where A, B are real and positive, and where a complex representation is made use of. Here \hat{i} and \hat{j} stand for the unit vectors along the positive x- and y-axes respectively. More generally, a wave with its electric vector oscillating along, say, the x-axis may correspond to a complex electric field intensity vector of the form $\hat{i}Ae^{i\delta}e^{-i\omega t}$, where δ ($0 \leq \delta < 2\pi$) is a constant phase.

In the following, the x-polarized component will always be taken to be of the form $\hat{i}Ae^{-i\omega t}$ (with A real and positive), while the y-polarized component will be taken to be of the form $\hat{j}Be^{i\delta}e^{-i\omega t}$ (with B , once again, real and positive) to indicate that there may, in general, exist a phase difference δ between the two components and that the real amplitudes of the two may differ.

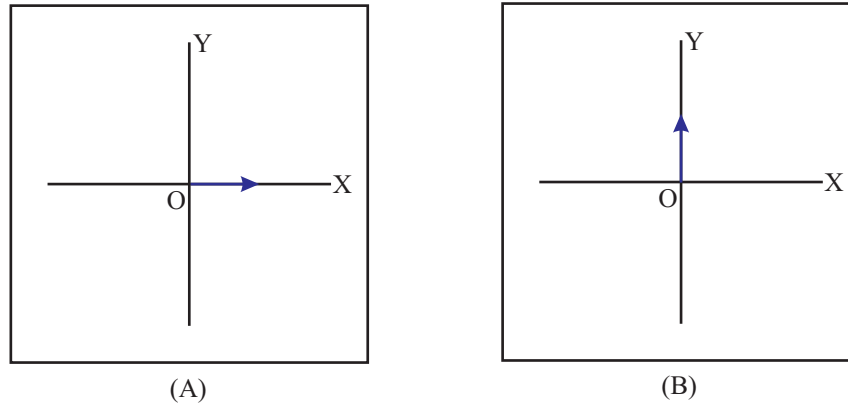


Figure 15-30: (A) x-polarized and (B) y-polarized waves; the direction of propagation is perpendicular to the plane of the figure, coming out of it; in (A), the electric vector oscillates along the x-axis of a Cartesian co-ordinate system, while in (B) it oscillates along the y-axis.

With this specification of the x-polarized and the y-polarized waves, let us recall how the various different states of polarization can be interpreted as superpositions of these two, with appropriate values of the phase difference δ and of the real amplitudes A and B .

15.5.3 Correlations: polarized and unpolarized light

First of all, the distinction between *polarized* and *unpolarized* waves can be stated as follows: *specific and definite values of A , B , and δ* correspond to the superposed wave being in some definite state of polarization. On the other hand, consider a *mixture* of superposed waves, where the two constituents of the mixture correspond to *uncorrelated* and arbitrary values of δ , and possibly of A and B as well. Such a mixture of superposed waves, with A , B , and δ being *random* variables, corresponds to an *unpolarized* wave. A *partially polarized* wave is one that can be looked upon as a mixture of polarized and unpolarized components.

15.5.4 Elliptically polarized light

Turning our attention now to a polarized wave with some definite values of A , B , and δ , the most general state of polarization is an *elliptic* one. One can explain this term as follows. The real part of the complex electric vector for the x-polarized and y-polarized waves at any given point in space (which, for the sake of concreteness, we take to be $z = 0$) are respectively $\mathbf{E}^{(x)} = \hat{i}A \cos \omega t$ and $\mathbf{E}^{(y)} = \hat{j}B \cos(\omega t + \delta)$ (recall that, in making use of the complex representation, one has to look at the real part of the complex wave function so as to recover the physically meaningful field variable (alternatively, one may consistently choose the imaginary part as well)).

When the two waves are superposed, the tip of the directed line segment representing the instantaneous electric vector of the resulting wave describes, in general, an ellipse in the x-y plane. Looking at the ellipse from the direction of the positive z-axis (i.e., with the light coming toward the observer), if the ellipse is described in a *clockwise* sense then the optical wave under consideration is referred to as a *right-handed* elliptically polarized one, while an opposite direction of rotation of the electric vector corresponds to *left-handed* elliptic polarization. States of left-handed and right-handed elliptic polarization are depicted in fig. 15-31.

One could equally well follow a different convention, in which case the terms left-

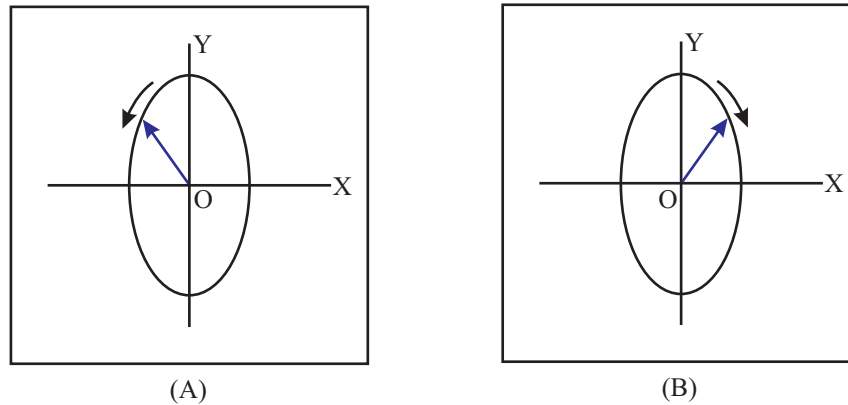


Figure 15-31: (A) Left-handed and (B) right-handed elliptic polarization; the tip of the electric vector describes an ellipse in the x-y plane, with the direction of describing the ellipse being different in (A) as compared to (B); the direction of propagation in either case is perpendicular to the plane of the figure, coming out of it.

handed and right-handed would get interchanged.

15.5.5 Circularly polarized and linearly polarized light

For the special case of the two constituent waves being of the same amplitude ($A = B$) and the phase difference being either $\frac{\pi}{2}$ or $\frac{3\pi}{2}$, the tip of the electric vector describes a *circle*, and one then has a *right-handed* ($\delta = \frac{\pi}{2}$) or a *left-handed* ($\delta = \frac{3\pi}{2}$) *circularly polarized* wave (see fig. 15-32(A), (B)).

Another special case corresponds to the phase difference being either 0 or π , for which the electric vector oscillates along a fixed straight line in the x-y plane, thus corresponding to a *linearly* polarized or *plane polarized* wave. The plane containing the line of oscillation and the direction of propagation is referred to as the *plane of polarization* (see fig. 15-33).

It is straightforward to verify that, for $\delta = 0$, the line of oscillation of the electric vector makes an angle θ with the x-axis where $\tan\theta = \frac{B}{A}$ (i.e., $0 \leq \theta \leq \frac{\pi}{2}$), as shown in fig. 15-33 (recall that we have assumed both A and B to be real and positive). On the other hand, with $\delta = \pi$, you can check that the angle is $\pi - \theta$, where θ is given by the same expression as above.

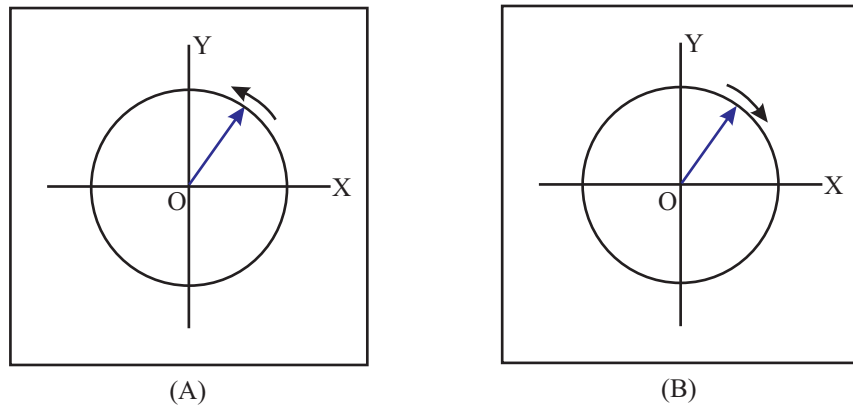


Figure 15-32: (A) Left-handed and (B) right-handed circular polarization; the tip of the electric vector describes a circle in the x-y plane, where the wave propagates along the z-direction, coming out of the plane of the paper.

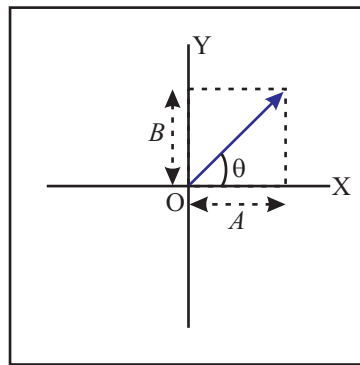


Figure 15-33: Linear polarization resulting from superposition of x- and y-polarized light; the electric vector oscillates along a line inclined at an angle θ to the x-axis; the value of θ (as also of the amplitude of oscillation) depends on the real amplitudes A and B of the two basic components, being given by $\theta = \arctan \frac{B}{A}$ when $\delta = 0$ (recall that the phase difference δ between the two constituents can be 0 or π for the superposed wave to be a linearly polarized one); the direction of propagation is perpendicular to the plane of the figure, coming out of it; the plane containing the direction of propagation and the direction of oscillation of the electric vector is referred to as the plane of polarization.

In summary, all the various states of polarization can be looked upon as superpositions of the two basic building blocks - the x-polarized and the y-polarized components, where the relevant parameters are the real amplitudes A , and B , and the phase difference δ . Unpolarized light corresponds to the situation where these three are effectively random variables.

15.5.6 Intensity relations

Considering the x-polarized component separately, without regard to the y-polarized component, its intensity is given by $I^{(x)} = |A|^2$ (up to a multiplicative constant, which we set equal to unity for the sake of simplicity), provided the real amplitude A has a definite, specific value. More generally, the intensity can be written in the form $I^{(x)} = \langle |A|^2 \rangle$, where the symbol $\langle \cdot \rangle$ denotes the *average value* of a quantity. Thus, if A is a random variable (in general complex), then $\langle |A|^2 \rangle$ denotes the average value of $|A|^2$, while if A is a deterministic variable, then it stands for the value of $|A|^2$ itself. In a similar manner, the intensity of the y-polarized component is given by $I^{(y)} = \langle |B|^2 \rangle$.

The intensity of polarized or unpolarized light obtained by the superposition of these two components is then given by

$$I = I^{(x)} + I^{(y)} = \langle |A|^2 \rangle + \langle |B|^2 \rangle. \quad (15-43)$$

In the case of polarized or partially polarized light, $I^{(x)}$ and $I^{(y)}$ may have different values while, for unpolarized light, one has

$$I^{(x)} = I^{(y)} = \frac{I}{2}. \quad (15-44)$$

Problem 15-13

Show that, for circularly polarized light,

$$I^{(x)} = I^{(y)} = \frac{I}{2}.$$

Answer to Problem 15-13

HINT: For circularly polarized light, one has $B = Ae^{i\delta}$, with $\delta = \frac{\pi}{2}$ or $\frac{3\pi}{2}$, where both A and B are real and positive, and are deterministic variables. Thus, $I_x = I_y = A^2 (= B^2) = \frac{1}{2}(I_x + I_y) = \frac{I}{2}$.

Note that this relation also holds for unpolarized light though, in the latter case, A and B are random variables (satisfying $\langle |A|^2 \rangle = \langle |B|^2 \rangle$; here again, one can choose A and B to be real).

15.5.7 Optical anisotropy: double refraction

For an optical wave in any given state of polarization, one can choose any two mutually perpendicular axes (in a plane perpendicular to the direction of propagation) as the x- and y-axes, and represent it as a superposition of an x-polarized and a y-polarized wave, with the parameters A , B , and δ represented by random variables, if necessary. However, such an arbitrary choice of the x- and the y-axes is allowed only if the medium under consideration is an *isotropic* one, i.e., its optical properties are *direction-independent*. For such a medium, the velocity of propagation of a linearly polarized wave is the same for all directions of polarization.

A medium may, on the other hand, be *anisotropic* in nature where, for a given direction of propagation, there exist *two specific directions* that prove to be appropriate choices for decomposing the wave into x- and y-polarized components (i.e., components whose superposition results in the wave under consideration). These two linearly polarized waves proceed in the medium along the given direction of propagation *with unequal velocities*. As a result, when an optical wave is incident on an interface at which it undergoes refraction and enters into an anisotropic medium, it *breaks up* into two linearly polarized components which proceed independently through the medium with two unequal velocities and, in general, along two different directions. This phenomenon is referred to as *double refraction*.

Certain crystalline substances like calcite, quartz, and tourmaline, are examples of optically anisotropic media.

15.5.8 Production of polarized light

When unpolarized light is incident on an interface separating two media, at a particular angle of incidence termed the *Brewster angle*, the reflected light is found to be linearly polarized. This is one way how linearly polarized light can be produced from unpolarized light emitted from a source.

Another widely adopted approach to produce linearly polarized light is to make use of a *polaroid*. One variety of polaroids is made of a thin sheet of polyvinyl alcohol (PVA) impregnated with iodine, the former being a polymer comprising of long-chained molecules. The sheet is produced in such a way that all the long molecular chains are aligned along one particular direction. By virtue of the doping with iodine, the chains acquire an electrical conductivity. Directions perpendicular and parallel to the molecules in the plane of the sheet may be thought as the x- and y-directions with respect to light incident normally on the sheet. If this light is unpolarized, then its y-polarized component gets absorbed in the sheet. This is because the sheet acts like a conductor with respect to the y-polarized component, in which the electromagnetic field is quickly absorbed, the electromagnetic energy being converted into heat and subsequently dissipated from the sheet. The x-polarized component, on the other hand, passes through the sheet with little attenuation.

Finally, linearly polarized light can also be produced by making use of the phenomenon of double refraction in optically anisotropic crystalline media. .

Doubly refracting crystals are also used to produce circularly and elliptically polarized light. Incidentally, linearly polarized light, on being reflected from the surface of a *conducting material* becomes, in general, elliptically polarized.

While the light obtained from traditionally used sources like an incandescent lamp or a flame is, in general, unpolarized, the light from a *laser* source (see section 15.6) is usually a polarized one.

As we will see in section 15.8, when unpolarized light is *scattered* by one or more scatterers, the scattered light is, in general, *partially polarized*.

15.6 Lasers: coherent sources of light

The development of the *laser* (Light Amplification by Stimulated Emission of Radiation) source has revolutionized the science of optics and has, at the same time, made possible a technological revolution of sorts.

The principle underlying the production of coherent light with a laser relates to *quantum theory*, to which you will find a brief introduction in chapter 16. It involves the concept of *photons* as energy quanta of electromagnetic radiation, of *energy levels* of atoms and molecules (or even of assemblies of atoms in material media), and finally, of the *interaction* between photons and atoms or molecules.

An isolated atom may be in any one of a number of quantum mechanical *stationary* states where each stationary state is characterized by a definite value of its energy. The state with the lowest energy is termed the *ground* state while those with higher values of the energy are referred to as *excited* states of the atom. The energies of the various stationary states (or, in brief, simply, *states*) form a *discrete* set, successive energy values being separated by energy gaps. A molecule, made up of several atoms held together, possesses similar discrete energy levels while an assembly of atoms in, say, a crystalline solid, is also characterized by energy levels that may come close together and form almost continuous sets, or *energy bands*, where two successive bands may once again be separated by an energy gap.

An electromagnetic field may also be in any one of a large number of states where each such state may be described as a collection of a large number of *photons*. A photon is a quantum of energy of the electromagnetic field and possesses other characteristics of a particle such as an energy, a momentum, and an angular momentum.

15.6.1 Emission and absorption as quantum processes

The production of light from a source involves, in the ultimate analysis, the emission of photons from the atoms or molecules making up the material of the source which, in turn, signifies an *interaction* of the atoms or the molecules with photons of the electromagnetic field around these. It is because of such interactions, that an atom or a molecule makes a *transition* from a higher to a lower energy state, giving out a photon. Similarly, the absorption of light by a material corresponds to photons interacting with the atoms or molecules of the material, raising them from lower to higher energy levels.

In other words, from a fundamental point of view, the emission and absorption of light are quantum mechanical processes (refer to sec. 16.10.4) where photons are emitted from or absorbed by atoms or molecules.

When an atom makes a transition from a lower to a higher energy level, it absorbs a photon from which it receives the necessary energy. In other words, an absorption process necessarily requires the presence of a photon in order that it may occur, which is in contrast to an emission process since in an emission the atom gives away a part of its energy in the form of a photon in making a transition from a higher to a lower energy level. An emission process can occur *spontaneously*, as depicted schematically in fig. 15-34(A), where the atom interacts with the ubiquitous *vacuum field* (refer, once again, to sec. 16.10.4), a background field that, according to the quantum theory, is always present even without the presence of localized or remote sources producing it. What is interesting, however, is that an emission can *also* take place in the presence of a photon, where the latter, in a manner of speaking, *provokes* an excited atom to emit part of its energy in the form of a second photon without itself being absorbed in the process, as depicted in fig. 15-34(B).

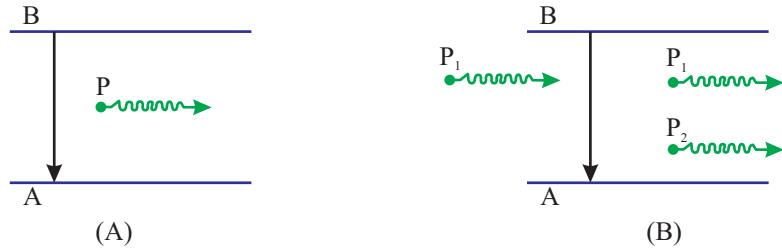


Figure 15-34: (A) spontaneous and (B) stimulated emission; A and B denote the ground state and excited state energies of an atom; in (A), the atom makes a transition from B to A, emitting a photon P, even without the presence of a second photon in its environment; in (B), on the other hand, a photon P_1 in the environment of the atom makes it likely that a second photon P_2 in the same quantum state is emitted when the atom makes a transition from the excited state B to the ground state A; this illustrative description, however, is not a precise one since the two photons, being identical particles, cannot be distinguished by labels like P_1 and P_2 ; one can only say that a single photon before the emission process has been replaced with two photons in the same quantum state after the emission.

15.6.2 The state of a photon

A photon, like any other particle, can be in any one of various possible states where the state of a photon is characterized by a wave vector (\mathbf{k}) and, correspondingly, an energy ($E = \hbar\omega = \hbar c|\mathbf{k}|$; we assume that the relevant medium is free space), and a certain *polarization*.

In the above expression for the energy of the photon, c stands for the velocity of light in vacuum, and \hbar for $\frac{h}{2\pi}$, where h denotes the Planck constant.

The polarization state of the photon is specified by a pair of vectors (say, \mathbf{a} , \mathbf{b}), both perpendicular to the wave vector \mathbf{k} , with their magnitudes satisfying $|\mathbf{a}|^2 + |\mathbf{b}|^2 = 1$, along with a certain *phase angle* ϕ . For instance, if \mathbf{k} is along the z -axis of a right-handed Cartesian co-ordinate system, and $\mathbf{a} = \hat{i}$, the unit vector along the x -axis, and $\mathbf{b} = 0$, with any arbitrarily chosen value of ϕ , then one has a photon in a linearly polarized state along the x -axis or, in brief, a x -polarized photon. Similarly, with $\mathbf{a} = 0$, $\mathbf{b} = \hat{j}$, the unit vector along the y -axis, with ϕ having any arbitrarily chosen value, then one has a y -polarized photon. Other choices for \mathbf{a} , \mathbf{b} , and ϕ correspond to *circularly* or *elliptically* polarized states of the photon.

Note the correspondence between this description of the state of polarization of a photon, and the polarization of a plane monochromatic wave outlined in sec. 15.5.2

15.6.3 Classical and quantum descriptions of the field

When one speaks of a plane progressive electromagnetic *wave* with a certain wave vector \mathbf{k} , where the wave is in a certain state of polarization, one is actually giving a *classical* description of the electromagnetic field. The description in terms of photons, on the other hand, is a quantum mechanical one. In the quantum description, a state of the electromagnetic field with a *large* value of the mean number of photons (and a correspondingly large statistical spread of this number), all belonging to the *same* state (or states that can in some sense be considered to be close to one another) is one that can, in a certain approximate sense, be described in classical terms. Here one speaks of the *mean* number of photons rather than some specific number since, in the quantum description, the values of various physical quantities appear as *random* variables that can only be specified in statistical terms. In such a statistical description, the mean value of a variable, as also its statistical spread, are of physical relevance.

In operational terms, this means that a large number of measurements of the photon number, all performed with the field in the same state, will yield results differing from one another with their mean value determined by the state of the field under consideration. The way the individual results are distributed or spread around this mean value is also determined by the state.

In other words, a classical description of an electromagnetic field as, say, a plane monochromatic wave with some definite state of polarization, can be said to be a valid one, if the quantum mechanical state of the field, described in terms of the states of the photons making up the field, is of a certain type, namely, one involving photons all in the same quantum state, and with a certain statistical distribution of the number of photons in this state. In such a correspondence between the classical and quantum descriptions of the electromagnetic field, the mean number of photons corresponds to the classical amplitude of the field, where the latter determines the field *intensity*.

15.6.4 Stimulated emission of radiation

Let us now consider an atom in an excited state placed in an electromagnetic environment such that, in the quantum description, there are N number of photons in a certain state (call it S), where N stands for the mean number and where an associated statistical distribution of the number is implied. For the sake of generality, one can assume that the field contains photons in *other* states as well. Suppose that the mean number (say, N') of photons in some other state (say, S') is less than the mean number in the state S .

Under these conditions, if the atom makes a transition to a state of a lower energy (say, the ground state), what will be the likely state of the photon that will be emitted in the process?

In the quantum mechanical description, every physical process has a certain *probability* associated with it, which gives the likelihood of its occurrence in comparison to other possible processes that may occur under given conditions.

The probabilities for the photon to be emitted in various possible states can be worked out in accordance with the basic rules of quantum theory. Following these rules, one finds that the probability of the photon being emitted in the state S is *larger* than that for the state S' because of N being greater than N' . This, then, is the basic fact characterizing stimulated emission: the greater the number of photons in a certain state in the electromagnetic environment of an excited atom, the greater is the probability of emission of a photon in that same state.

Evidently, such a process of stimulated emission can end up being a *snowballing* one. For, suppose that one has an assembly of atoms in the excited state in an electromagnetic environment in which photons in some particular state S are more numerous compared to those in other states (like S' above). Then each atom will emit a photon preferentially in the state S , leading to an *increase* in the mean number N of photons in this state, as a result of which the next emission from some other atom in the assembly

will be even more likely to produce a photon in the same state S .

In other words, the process of stimulated emission of radiation, which occurs with atoms in their excited states placed in an appropriate electromagnetic environment is likely to produce a field with a very large number of photons, all in the same state. The associated statistical distribution of the number of photons then approaches a certain standard form, and the electromagnetic field can then be described in classical terms. One can choose the state S in such a way that the resulting classical field corresponds to a plane monochromatic wave characterized by a well defined wave vector, and in some specific state of polarization. What is more, the wave can be described by a certain specific value of the *phase*.

1. In a quantum mechanical state of an electromagnetic field characterized by photons distributed over their possible individual states, the phase is often not a meaningful concept.
2. The concept of phase becomes meaningful in certain states of the electromagnetic field including the ones that can be described in classical terms as monochromatic plane waves with well-defined states of polarization.
3. While talking of the phase, one means only the constant part of the total phase of a wave. the latter includes space-and time dependent parts as well.
4. Even for a classical monochromatic plane wave, the phase is of physical relevance only with reference to the phase of some other wave. Considering a wave all by itself, one can choose its phase ϕ to be zero.

15.6.5 Stimulated emission and coherent waves

The idea of coherence has been referred to several times in chapters 9 and 14, as also in the present one (see, for instance, sec. 15.3).

A plane wave with well-defined, definite values of the frequency (ω), wave vector (\mathbf{k}), amplitude (A), and phase (ϕ), and in a well-defined state of polarization, is said to be a *coherent* one. One can, on the other hand, have *incoherent* waves which can be described as *statistical mixtures* of waves with various different values of ω , \mathbf{k} , A , ϕ , and the state of

polarization. Depending on the extent of the admixture, one can have waves of various different *degrees* of coherence, where one can express the degree of coherence in terms of a number of quantitative measures.

1. While I have illustrated the idea of coherence here by referring to plane waves, the concept of coherence works equally well for other types of waves such as spherical or cylindrical ones.
2. In the context of interference of waves, I have talked of coherence of a pair of waves that get superposed with each other in a given region of space. The idea of coherence between two waves is not basically different from that of a single wave. Consider, for instance, the superposed wave resulting from two or more waves. If this superposed wave happens to be a coherent one then the individual waves under consideration are said to be coherent with respect to one another. While it is relatively simple to express the concept of coherence in quantitative terms for a single plane or spherical wave, quantitative expressions relating to the degree of coherence of waves of more general description involve more detailed considerations.

Commonly used sources of light produce incoherent waves or, at most, partially coherent waves with comparatively small values of the degree of coherence. A laser source, on the other hand, produces coherent waves by making use of stimulated emission of photons by excited atoms of a material placed in an electromagnetic environment that contains an initial concentration of photons in certain states to start with where there is a mechanism of *selecting out* photons in a specified state. This results in the production of a coherent wave that can be described classically in terms of well-defined and specific values of ω , A , k , and ϕ , and a specific state of polarization.

Even starting from an unpolarized wave, a wave with a specified state of polarization can be produced by one of several means, say, by making the wave pass through a *doubly refracting crystal*.

15.6.6 Population inversion

As mentioned above, the production of a coherent wave of light requires a continual increase, or *amplification*, in the mean number of photons in a certain specific state, commonly referred to as a *mode*, by means of stimulated emission of photons in that mode from an assembly of excited atoms or molecules. More precisely, let the atoms in a given assembly be in any of two possible states, namely, a ground state and an excited state, where a fraction x of the atoms is in the excited state, with the remaining fraction $1 - x$ being in the ground state. Then an amplification of the mean number of photons in a given mode requires that the value x be larger than 0.5, since otherwise there will be not enough atoms in the excited state to add to the number of photons in a single mode by stimulated emission.

What this implies is that an *amplification by stimulated emission cannot be effected with an assembly of atoms in a state of thermal equilibrium* since, in an assembly in thermal equilibrium at any given temperature, the mean number in a higher energy state is always *less than* that in a lower energy state. This is a result of quite general validity, being a consequence of the *Boltzmann distribution* formula characterizing the relative probabilities of states of a system in thermal equilibrium at any given temperature (see sec. 8.14).

One can, however, achieve light amplification by stimulated emission from an assembly of atoms maintained in a *steady state, away from an equilibrium state*, where a *population inversion* has been produced, i.e., the fraction of atoms in the excited state has been made larger than that in the ground state. This requires, in particular, an assembly of atoms with at least *three* relevant energy levels with energies, say, E_0 (the ground state energy), E_1 (the energy of an excited state), and E_2 (the energy of a *second* excited state).

An electromagnetic field of appropriate frequency is used to raise the atoms from the ground state to the excited state of energy E_1 by the process of absorption, where E_1 is larger than E_2 . A number of the excited atoms with energy E_1 then radiate energy by spontaneous emission and make a transition to the excited state with energy E_2 , where

the latter is chosen to be a *metastable state*, i.e., one which takes a relatively long time to make a transition down to the ground state with energy E_0 because of some *selection rule* or other, due to which the probability of the transition has a relatively low value. In consequence, there occurs an accumulation of atoms in the metastable state, whose population becomes larger than that of the ground state. There now takes place an amplification of photon number by stimulated emission from this metastable state to the ground state. Such a scheme of the energy levels (the ‘three-level scheme’) and the transitions involved is depicted in fig. 15-35.

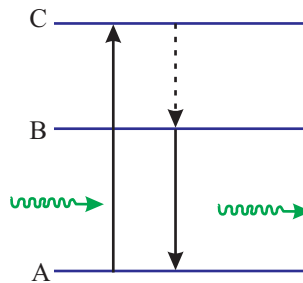


Figure 15-35: The three-level scheme; A, B, and C denote the ground state and two excited states of an atom; a photon, denoted by the wavy line on the left, is absorbed by the atom in the ground state, which makes a transition to the upper excited state C; the atom then makes a transition to the intermediate state B, which is a metastable one; the transition is shown with a dotted line since it is usually a *non-radiative* one, where the energy is given out in the form of internal energy of an assembly of atoms rather than in the form of a photon; finally, the atom makes a transition from B to A by means of stimulated emission; the emitted photon is shown by the wavy line on the right; the photons in the electromagnetic environment causing the stimulated emission are not shown.

1. In practice, the number of energy levels involved in the operation of a laser is usually greater than three.
2. The electromagnetic field made up of the photons absorbed by the atoms in the ground state to make a transition to the excited stat corresponding to the energy level C in fig. 15-35 is referred to as the *pumping* field.
3. In atomic physics, not all possible transitions from higher to lower energy levels of an atom are found to occur in reality. The probabilities of some of the transitions are so low that these transitions are not observed in practice. Such transitions are referred to as *forbidden* ones in contrast to *allowed transitions* which are commonly observed to occur. The rules of quantum theory can be invoked to distin-

guish between allowed and forbidden transitions. However, the term ‘forbidden’ is not to be taken in an absolute sense since a forbidden transition often does occur in reality, though with a small value of the probability of transition. An excited state from which a transition to a lower energy state occurs with such a low probability is termed a metastable state since the atom can continue to be in this state for a considerably long time.

4. In quantum theory, a state of a system with a given energy is characterized by one or more *quantum numbers* (see chapters 16, 18). In calculating the probabilities of transition between various states, one encounters certain *selection rules* involving these quantum numbers that determine whether a given transition is allowed or forbidden. For instance, if a quantum number has values n_1 and n_2 for the initial and final states of a transition, then a condition of the form $n_2 = n_1 \pm 1$ may indicate that the transition is an allowed one, a violation of this relation then corresponding to a forbidden transition.

15.6.7 Light amplification in a resonant cavity

Just achieving a population inversion is not the end of the story, since one has to *maintain* this population inversion so as to obtain a coherent wave of sufficient intensity for a sufficiently long duration of time. In other words, the assembly of atoms is to be maintained in a steady state away from thermal equilibrium (states with population numbers *oscillating* in time are also useful in the production of pulses of coherent light). This requires that photons produced in stimulated emission in a given quantum state are *not to escape* from the environment in which the assembly of atoms is kept. Such an environment of a high concentration of photons in a specified state, resulting in a continuing condition of population inversion in the assembly of atoms, is made possible by a *resonant optical cavity*.

In practice, the resonant cavity may consist of a pair of highly reflecting surfaces parallel to each other (fig. 15-36), kept at an appropriate separation. Such a cavity selects out photons with a wave vector k perpendicular to the reflecting surfaces, the magnitude of the wave vector (and, correspondingly, the frequency of the photons) being determined

by the separation between the surfaces. Only the photons in this particular quantum state can stay for a long time inside the cavity, being reflected repeatedly between the two surfaces, while photons in other states quickly leak away from it. In the classical description, a *stationary wave* with a large amplitude, characterized by an wave vector k (or, more precisely, by the pair of wave vectors k and $-k$) is set up in the cavity. A propagating wave with a wave vector k can then be taken out of the cavity for any required duration of time by operating an *optical shutter*, whose principle I do not enter into here. The wave coming out of the cavity is commonly in the form an intense and narrow beam of coherent light.

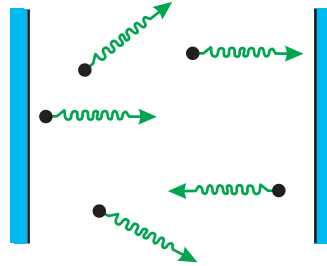


Figure 15-36: Illustrating the idea of an optical resonator; the resonator essentially consists of the space between a pair of parallel reflecting surfaces; photons in a certain specific quantum state are reflected back and forth between the plates; photons with even a slightly different value of the frequency or wave vector are either attenuated or leak away from the region, as shown by the slanting wavy lines; as the number of photons gets amplified by stimulated emission, an electromagnetic field admitting of a classical description as a stationary wave builds up in the cavity; a laser beam in the form of a progressive wave can be taken out by activating an optical shutter (not shown in the figure).

The initial build-up of photons in the cavity, necessary for the stimulated emission and light amplification to take place, occurs by means of a process of a statistical nature, and an externally applied field made up of such photons is not essential for this.

15.6.8 The laser as a coherent source of light: summary

Electromagnetic waves in general, and light waves (i.e., waves in the optical range of frequencies) in particular, are characterized by their degree of coherence, where a coherent wave is described by well-defined values of frequency, wave vector, amplitude,

and phase. An incoherent wave, on the other hand, is made up of an uncorrelated admixture characterizing these parameters. Coherent waves give rise to interference and diffraction patterns consisting of alterations of intensity, of a regular nature, in space.

A laser is a coherent source of light where the phenomenon of stimulated emission by atoms and molecules is made use of. Emission and absorption of light by atoms is, from the fundamental point of view, a quantum phenomenon where the interaction of an electromagnetic field with matter is described in terms of the interaction of photons with the atoms. In particular, stimulated emission of radiation is a typically quantum phenomenon, in which the presence of photons in any given quantum state in the environment of the emitting atom increases the probability of emission of a photon in the same state.

When stimulated emission takes place from an assembly of atoms kept in an optical resonator, and each of the atoms possesses at least three energy levels involved in a scheme as shown in fig. 15-35, light *amplification* may be made to take place by means of stimulated emission of radiation from these atoms. Of the three levels shown in fig. 15-35, the one with the intermediate value of energy has to be a metastable one so that population inversion may be achieved, which is a necessary condition for light amplification.

The *laser diode*, which is a practical device for the production of a laser beam, is briefly introduced in sec. 19.3.7.3.

15.7 Holography

The development of the laser as a source of coherent light has made possible, among a multitude of quite remarkable developments, the practical realization of *holography*, the technique of recording the details of an object in a hologram in the form of *phase information*, and of subsequently producing a three dimensional image of the object from the phase information, even without the presence of the object itself.

The way a hologram of an object is created and then used to view the image of the object is illustrated in fig. 15-37(A) and (B). A coherent plane monochromatic wave is directed to the object O with a photographic film P placed behind the object. The wave incident on P results from a superposition of two waves, of which one is the *reference* wave that may be made to be incident on P after being reflected from a mirror M, and the other is the *object* wave sent out to P from various different parts of the object as the wave incident on it interacts with these parts and gets modified by such interaction. The modified wave arriving at any given point (A) of P from any point on the object (say, B) can be represented in terms of an amplitude and a phase, where the latter depends on the optical path length of the ray path extending from B to A. Of the two, it is the phase that is of greater relative importance in creating the record on P.

Considering a typical point, say, A on P, the intensity at A resulting from the superposition of the reference wave and the object wave (the two being commonly generated from the same coherent source) depends on the phase difference of the two, where the object wave is made up of contributions from waves sent out from all the various points like B on O. In other words, the record in even a tiny region of P involves amplitude and phase information from *all* the points of O resulting from the superposition of the object wave with the reference wave. The hologram is obtained by developing the photographic film and taking a positive print on a second transparent film.

Once the hologram is created with all this record of the details coming from the various different points of the object, a three dimensional image, with all these details re-created, is obtained by illuminating the hologram with light from the same coherent source (commonly, a laser) as was used to generate the reference wave and the object wave for producing the hologram. This *reconstruction* wave (terms such as reconstruction *beam*, object *beam* are also in common use), interacts with the hologram and produces diffracted waves as in fig. 15-37(B) forming two three dimensional patterns, both of which are the reconstructed images of the object. Of these, one is a virtual image, points on which are obtained by producing the diffracted ray paths backwards, while the other is a real image, where ray paths intersect. In other words diffracted waves from the hologram appear to diverge from points on the virtual image, while they

converge to (and thereafter diverge from) points on the real image.

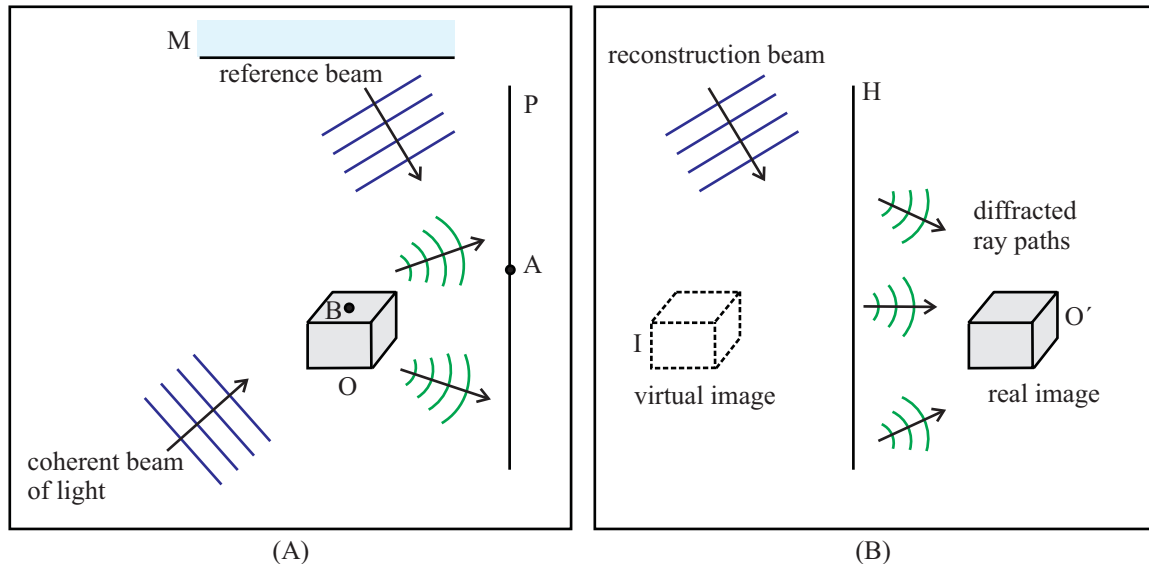


Figure 15-37: (A) the production of a hologram; the object O is illuminated by a coherent wave; the superposition of the object wave and the reference wave creates a record of phase information on the photographic plate P ; the wave scattered from any point B on O reaches the photographic plate P , and interferes with the reference wave, producing a resultant intensity at any point like A on P determined by the phase difference between the two; considering all such points on O , a complex phase record is created on P ; (B) image reconstruction; as the reconstruction wave is diffracted by the hologram H (generated from P) to the right, a real image is formed out of the diffraction pattern; at the same time, a virtual image is created on the other side of H .

The basics of holography can be grasped by considering the hologram of a *point object*. Fig. 15-38(A) shows a coherent plane wave incident on a point object O , with a photographic plate P behind it. Looking at any point A on the plate, it is reached by both the direct wave incident on the plate (the reference wave in the present context) and the wave scattered from O , which in this instance constitutes the object wave. The two waves get superposed and interfere, and the intensity at A is determined by the phase difference between the two (recall from sec. 15.3.1 that variation in the amplitude of the waves with the position of the point A can be ignored since this variation is a slow one compared to the variation in the phase). The latter, in turn, is determined by the optical path difference which, in the present case, is given by $(OA-BA)$, i.e., by $(OA-OC)$ in the figure. Denoting the distances OC and CA by d and r respectively, and considering

values of r small compared to d for the sake of simplicity, the phase difference is seen to be

$$\delta = \frac{2\pi}{\lambda} \frac{r^2}{2d}, \quad (15-45)$$

where λ stands for the wavelength of the coherent wave used for creating the hologram.

Thus, there occurs a maximum or minimum of the intensity at the point A if its distance from C satisfies

$$r = \sqrt{2d\lambda n}, \text{ or } \sqrt{2d\lambda(n + \frac{1}{2})}, \quad (n = 0, 1, 2, \dots), \quad (15-46)$$

respectively. Evidently, the points on the photographic plate corresponding to maxima and minima of intensity, as determined by the phase difference δ , lie on circles with their centers at C. On developing the photographic plate and taking a positive print on a transparent film, one ends up with what is referred to as a *zone plate*, where there are alternating dark and transparent annular zones (fig. 15-38(B)), with the radii of the latter given by $\sqrt{2d\lambda n}$, ($n = 0, 1, 2, \dots$).

Suppose now that the zone plate, which is the hologram of the point object O, is illuminated with the reconstruction wave, which is once again a coherent plane wave with wavelength λ , as in fig. 15-39, but with the point object O now absent. The transparent annular zones now act as circular slits diffracting the wave. The intensity at any point due to the diffracted wave can be obtained by considering the waves reaching that point from the various annular zones and working out the expression for the superposed wave. Such an exercise shows that the intensity varies with the position of the point under consideration, with a *maximum* of intensity occurring at certain points. The positions of these points can be easily worked out by once again considering the phase difference of the waves reaching a point from the successive annular zones. In particular, the intensity at a point for which the phase difference is 2π will be a maximum, since the waves reaching this point from the successive zones all interfere constructively.

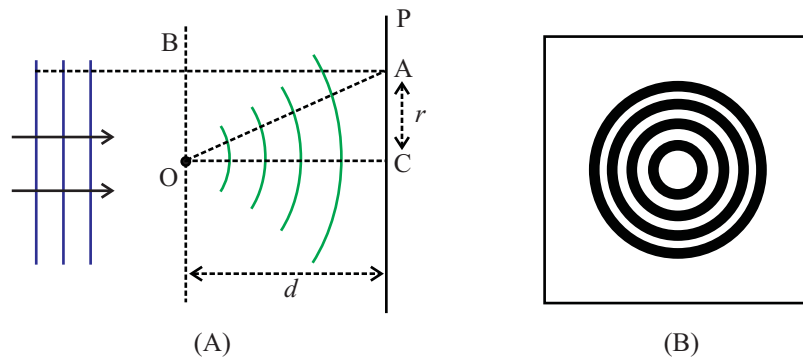


Figure 15-38: (A) Producing the hologram of a point object O ; on being illuminated with an incident plane wave, the point object produces a spherical scattered wave which interferes on the photographic plate P with the reference wave which may be the same wave as the incident one; the resulting intensity at any point A on P depends on the path difference $(OA-OC)$ where C is the point for which the path difference is zero; in equations (15-45), (15-46), we assume $d \gg r$; (B) the hologram, consisting of transparent concentric annular regions with radii proportional to the square roots of positive integers, interspersed with dark regions; the radius of a transparent zone corresponds to the mean of the inner and outer radii.

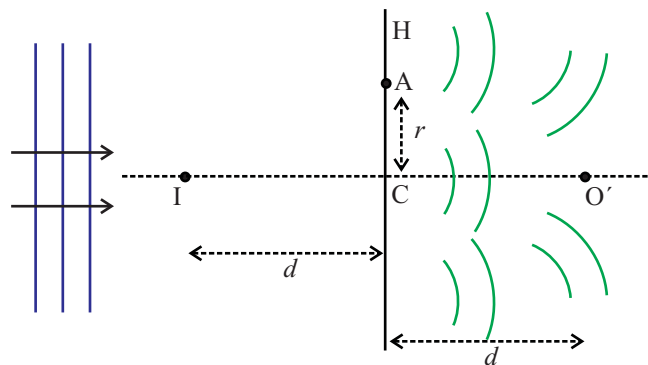


Figure 15-39: Image reconstruction from the hologram of a point object; the distances of O' and I from the hologram H is the same as that of the object point O from P in fig. 15-38(A); the point A is in the n th zone, whose radius is given by the first expression in (15-46); diffracted waves from successive zones interfere constructively at O' , which is the real 'image'; I is the virtual image that can be viewed from the other side (i.e., from the right) of the hologram H .

The point O' in fig. 15-39, situated to the right of the hologram on the line OC produced such that $OC=OC'$ is precisely such a point (recall that the point object O was located to the left of the photographic plate at the time of taking the hologram; this point is not shown in the figure since the object is now absent; the previous location of O is now denoted by I , see below) at a distance d from the center C of the zone plate.

Problem 15-14

Referring to fig. 15-39, consider points A and A' on the zone plate at distances $\sqrt{2d\lambda n}$ and $\sqrt{2d\lambda(n+1)}$, where n is any positive integer (the point A' is not marked in the figure). Show that $AO'-A'O' = \lambda$.

Answer to Problem 15-14

Note that the point I in fig. 15-39 is located at the same position with respect to the plane of the hologram as O in fig. 15-38(A) (see below), i.e., in the recording phase, and hence $AO'-A'O'=AO-A'O=\lambda$, by the first relation in (15-46), since by the construction of the half period zones, A and A' are located on the n th and $(n+1)$ th bright rings (maximum intensity, i.e., blackened regions on the photographic plate, giving rise to transparent regions in the hologram, obtained as the positive print on developing the plate). More explicitly, $AO' = (d^2 + 2dn\lambda)^{\frac{1}{2}} \approx d + n\lambda$ (for $n\lambda \ll d$), and $A'O' \approx d + (n+1)\lambda$, which gives the required result.

NOTE: This shows that the phase difference between the waves reaching O' from successive zones is 2π , i.e., these waves interfere constructively at O', resulting in an intensity maximum at this point.

This point (O') is referred to as the *real image* in the holographic reconstruction. Along with it, there occurs a *virtual image* I on the other side of the hologram, i.e., to the left of the zone plate in fig. 15-39. While the diffracted waves sent out by the zone plate to the region on its right actually reach the point O', they appear to diverge from I where the diffracted waves from the successive zones, appearing to have originated at I, once again have a phase difference of 2π since, as shown in the figure, $IC=O'C = d$. An observer looking from the right of the zone plate can observe the virtual image as a reconstruction of the object O used to produce the hologram.

1. In reality, O' and I are not images of O in the usual sense of the term, but are diffraction maxima produced by the hologram. Indeed, the object O is not even present when the diffraction maxima are produced. These diffraction maxima are nothing but a reconstruction of the object from the phase information recorded in the hologram when the latter was produced by interference between the object beam and the reference beam.
2. Other diffraction maxima are also produced on the line IO' shown in fig. 15-39, for

which the phase differences of the waves from the successive zones are $4\pi, 6\pi, \dots$

These secondary maxima are not of relevance in holographic reconstruction since they are of much lower intensity.

The principles underlying the holographic reconstruction of an extended object are similar, since each point of the object gives rise to a phase recording on the hologram in an analogous manner. On illuminating the hologram with the reconstruction beam, a real and a virtual ‘image’ are formed. With the observer located in the diffraction region (to the right of the hologram in the case of fig. 15-37(B)), the virtual object can be seen on the other side of the hologram. The virtual holographic image captures all the three dimensional details of the object and can be viewed from different angles just as in the case of the original object.

15.8 Scattering of light

The term ‘scattering’ refers to the phenomenon of waves encountering *small* objects, like obstacles or apertures, and getting altered due to their interaction with these objects, where the latter are referred to as ‘scatterers’. What is relevant here is that the dimension of the scatterer is to be small so as to be comparable with (say, 10^{-2} to 10^2 times) the wavelength (λ) of the wave. Confining our attention to the domain of optics this means, roughly, a linear dimension in the range 10^{-8} m to 10^{-4} m. Moreover, with scatterers of this dimension, it is often necessary to look into the interaction of an electromagnetic wave with not one single scatterer, but with a *collection* of these. One requires special methods to describe the way the electromagnetic wave gets modified, first, by a single scatterer of linear dimension lying in the above range, and then by a given collection of scatterers.

15.8.1 Rayleigh scattering

If the linear dimension of the scatterer be of the order of the wavelength or a fraction thereof, one speaks of *Rayleigh scattering* since Lord Rayleigh was the first to put forward a theory of such a scattering process while explaining the blue of the sky and

several other optical phenomena.

What essentially happens in Rayleigh scattering is that the oscillating electric field of the wave incident on the scatterer modifies the states of motion in the electrons in its atoms or molecules, initiating forced oscillations of these electrons at a frequency equaling that of the incident wave. These oscillating electrons give out a part of their energy in the form of electromagnetic radiation, which appears as the scattered light. This can be described as *dipole radiation* from the atoms and molecules since the forced oscillations of the electrons set up harmonically varying electric dipole moments in these.

Since the energy necessary to set up the forced oscillations of the electrons in the first place comes from the incident wave itself, the process can be described as an absorption of part of the energy of the incident wave and re-emission of this energy in the form of the scattered wave.

The basic process underlying Rayleigh scattering is identical to the one responsible for *dispersion* in a material medium; indeed, the phenomenon of dispersion can be looked upon as one resulting from the coherent superposition of scattered waves from a homogeneous distribution of scatterers making up the medium under consideration.

15.8.1.1 Rayleigh scattering by a single scatterer

For the sake of concreteness, I will first describe the principal features of Rayleigh scattering by considering a linearly polarized incident wave, where the process is depicted schematically in Fig. 15-40. The direction of propagation of the incident wave is along the line OC, while OA is the line along which the electric vector of the incident wave oscillates and, consequently, the forced oscillation of the electron takes place. Choosing a spherical polar co-ordinate system with its polar axis along OA, one can work out the rate of emission of energy (say, P) per unit solid angle (see section 11.8.2.2 for an introduction to the concept of solid angle) in any given direction specified by a polar angle θ and an azimuthal angle ϕ .

A spherical polar co-ordinate system consists of a *polar axis* like OA and a plane (the

'plane of the azimuth') containing the polar axis which, in the present instance, is chosen to be the one perpendicular to the plane of OA and OC. The spherical polar co-ordinates of any point (say, P in fig. 15-40) are then the distance $r = OP$, the angle θ made by OP with OA, and the angle ϕ between the plane of OA and OP and the plane of the azimuth. In the figure, the point P has been chosen to lie in the plane of OA and OC, which implies $\phi = \frac{\pi}{2}$; such a special choice does not imply a loss of generality in this case since the scattering features are independent of ϕ because of the symmetry of the problem.

The figure shows a polar plot of P in which the distance of any point (say, P) from the origin gives the power radiated per unit solid angle for the corresponding value of θ and ϕ . The symmetry of the situation implies that the power in a given direction has to be independent of the azimuthal angle ϕ . Hence the figure shows only the variation with θ , for a fixed value of ϕ .

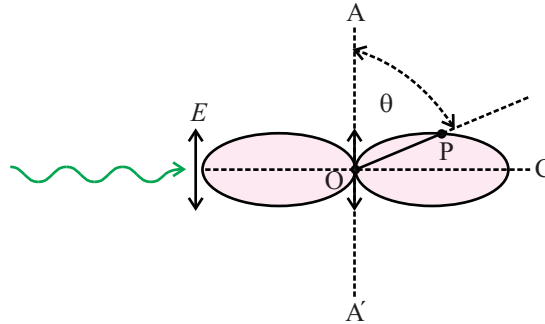


Figure 15-40: Illustrating Rayleigh scattering; OC is the direction of propagation of an incident polarized wave, with its electric vector oscillating parallel to OA; the latter specifies the direction along which the forced oscillations of the electrons are set up; OP is any chosen direction along which the scattered radiation is viewed; the polar angle made by OP with OA is θ ; since the radiation is emitted by the oscillating electrons after being absorbed from the incident radiation, the radiation is azimuthally symmetric about OA, and OP can be chosen to lie in the plane of OA and OC in working out the power (P) radiated per unit solid angle in the direction θ ; a logarithmic polar plot of $\log P$ against θ is shown; the scattering is maximum in the meridional plane $\theta = \frac{\pi}{2}$, and is symmetric in the forward and backward directions (a consequence of the azimuthal symmetry).

One observes from the graph that the power radiated along $\theta = 0$ is zero, and is maximum for $\theta = \frac{\pi}{2}$. As seen from the figure, the radiation is symmetric in the forward and

backward directions, and decreases monotonically as the transverse direction ($\theta = 0$) is approached.

In addition to the directional distribution of the scattered radiation, two other important characteristics of the scattering are the polarization properties of the scattered light, and its wavelength dependence. Assuming that the incident radiation is polarized as indicated above, the scattered light in any given direction is also found to be polarized, with the electric vector oscillating in the plane containing the line of sight (along a direction perpendicular to it) and the electric vector of the incident light, the latter being parallel to the line OA for the situation depicted in figure 15-40.

Finally, the total power radiated in Rayleigh scattering (as also the power per unit solid angle in any given direction) as a function of the wavelength (λ) of the incident radiation goes like λ^{-4} : the blue end of the visible spectrum gets scattered by as much as 10 times compared to the red end.

Corresponding statements hold for incident light polarized along the line OB (not shown in fig. 15-40) perpendicular to both OA and OC. The power radiated per unit solid angle will now be azimuthally symmetric about OB. The scattered light is, once again, linearly polarized, with its plane of polarization made up of the line of sight and the line OB.

What happens if the incident radiation, coming in along OC in fig. 15-40 is unpolarized? The λ^{-4} -dependence of the power radiated - the tell-tale feature of Rayleigh scattering, continues to hold. The power radiated per unit solid angle is now axially symmetric about the line OC, i.e., the direction of propagation of the incident radiation, and depends only on the angle between the line of sight and the direction OC. One can, without loss of generality, choose the line of sight in the plane of OC and OB shown in fig. 15-41 (the line OB is, however, not shown in fig. 15-40; one could equally have chosen the plane of OC and OA for this purpose), in which case a polar plot of the power radiated as a function of ϕ (the angle between the line of observation OP, and OC; this now becomes the polar angle of OP with reference to OC) looks as in the figure.

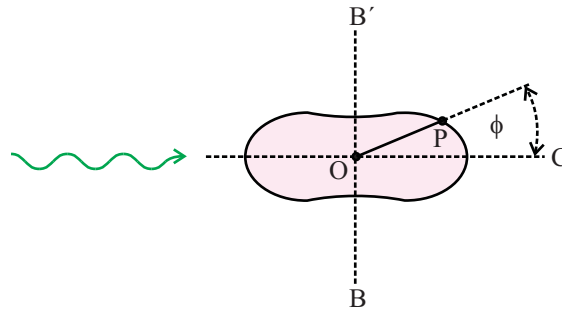


Figure 15-41: Rayleigh scattering with unpolarized incident radiation; with reference to fig. 15-40, the electric vector now has components along OA and OB (perpendicular to OA and OC, not shown in fig. 15-40); the scattering features for unpolarized radiation are now axially symmetric about OC, and a logarithmic plot of the power per unit solid angle against the angle (ϕ) between OC and OP is shown, with OP chosen to lie in the plane of OC and OB (with extension BOB'); the scattering is maximum along the forward and backward directions, but is non-zero along the transverse direction as well.

The power radiated is once again seen to be symmetric in the forward and backward directions, being a maximum in these two directions and a minimum, but non-zero, in the transverse directions. Finally, the scattered light is, in general, *partially polarized*. Considering a plane perpendicular to the line of sight, the scattered light can be looked upon as a superposition of two basic components, one with the electric vector perpendicular to the plane containing the direction of incidence and the line of sight, and the other with the electric vector lying in this plane. In longitudinal view (line of sight parallel to OC) the scattered light is unpolarized, while in transverse view (line of sight parallel to OB) it is linearly polarized (fig. 15-42). In between, the scattered light is partially polarized with the degree of polarization (a measure expressing how close the light is to a linearly polarized one) increasing monotonically.

1. In a quantum theoretic description (see chapter 16 for an introduction to the basics of quantum theory), Rayleigh scattering can be described as an elastic scattering of photons from the scatterer, say, an atom or a molecule. The scattered radiation results from the excitation and de-excitation of internal modes in the scatterer but, on the whole, there is negligible energy transfer to the internal modes. The scatterer as a whole takes up momentum from the photon, but the associated energy is negligibly small because of the large mass of the scat-

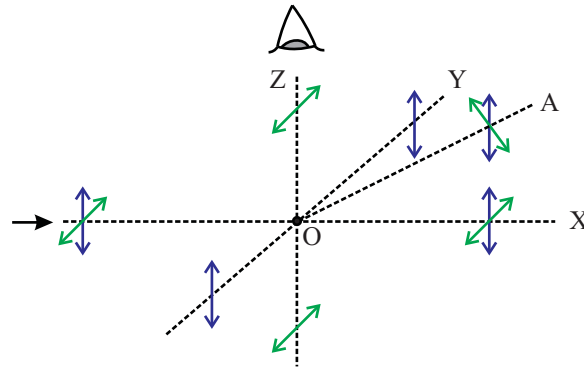


Figure 15-42: Polarization of scattered light in Rayleigh scattering with unpolarized light; the scattered light is unpolarized in longitudinal view (OX, direction of incidence) and linearly polarized in transverse view (OY, OZ); in between (say, along OA), it is partially polarized; note that the relevant directions have been renamed with reference to fig. 15-40 and 15-41 (OX in place of OC; OZ, OY in place of OA, OB; and OA in place of OP).

terer. Consequently, the change in frequency of the photon due to scattering is negligible.

2. There occurs transitions between states of the electrons in the scatterer which causes the emission of scattered radiation, but the process itself cannot be described as a scattering of photons by the electrons. The latter are *bound* in the scatterer, and it is actually the latter that scatters the radiation as a whole.

15.8.1.2 Rayleigh scattering in a fluid

Rayleigh scattering is commonly known by its explanation of the blue of the sky. As sunlight passes through the sky in reaching the earth, it gets scattered by molecules making up the gaseous atmosphere. The basic scattering process involved here is the scattering of light by the individual molecules of the atmosphere.

When light passes through a dilute gas, the waves scattered from individual molecules in any given direction are *uncorrelated with one another*, since the average separation between the molecules is large, as a result of which the path difference of waves scattered from individual molecules exceeds the coherence length of the radiation. In other words, the radiations scattered from the different molecules are *incoherent*, and the intensity of the scattered radiation in any given direction is obtained by summing the

intensities scattered by the individual molecules.

The degree to which Rayleigh scattering causes the incident energy of radiation to be diminished as the radiation travels through it is commonly expressed in terms of a quantity referred to as the *attenuation coefficient* which is directly related to the rate at which energy is removed from the incident beam and scattered away in all directions. For a dilute gas, one can calculate this from Rayleigh's formula for the total power scattered by an individual scatterer, and then making use of the incoherent nature of the scattered radiation. The result is found to agree well with experimental observations. The attenuation coefficient is found to depend on the wavelength through the same factor of λ^{-4} as it should. The angular distribution and the polarization properties of the scattered radiation remains essentially unchanged as compared to the corresponding features for a single scatterer. In addition, the attenuation coefficient is found to be inversely proportional to the number density of molecules in the gas.

The blue of the sky is explained by noting that, while looking at the sky, what we commonly observe is the sunlight scattered from the atmospheric layers, which is rich in the blue-violet component of the visible spectrum because of the λ^{-4} -dependence of the scattered energy.

Rayleigh scattering also occurs in *denser* gases and liquids or in media where dense aggregates of particles are suspended, as in a colloidal solution. The mean separation between the atoms or molecules, or between the suspended particles, and their mean free path are now comparable to the wavelength, in consequence of which appreciable *density fluctuations* occur over regions of space whose dimension is small compared to the wavelength. Rayleigh scattering then takes place from these density fluctuations that appear as tiny inhomogeneities in the medium. The scatterers being densely packed in this case, the scattered radiations from neighboring scatterers possess a considerable degree of *coherence* and there occurs a summation of amplitudes rather than of intensities. The resulting attenuation coefficient can be worked out, and is once again characterized by the λ^{-4} dependence on the wavelength of radiation. In the limit of low density of the scatterers, the dilute gas formula is recovered. In other words,

the scattering from individual molecules in a dilute gas and the scattering from density fluctuations are not fundamentally different processes, differing only in the degree of correlation between the radiation scattered from neighboring scatterers.

15.8.2 Mie scattering

If the scatterer be larger in size compared to the sub-wavelength scatterers responsible for Rayleigh scattering, being, say, several times the wavelength of light, a number of distinctive features are found to characterize the scattered radiation as compared to those in Rayleigh scattering. The scattering by such larger particles is commonly referred to as Mie scattering, since it was Mie (with independent contributions made by Debye) who put forward a complete theory of scattering of electromagnetic waves by a spherical particle of any given radius, where the particle may be either a conductor or a dielectric. Since the radius of the sphere in this theory can have any given value, one can consider special cases where the radius is small or large compared to the wavelength λ , or has an intermediate value comparable to λ . In the limit of small size of the scatterer, one actually recovers the results relating to Rayleigh scattering.

While Mie's theory gives precise results (in the form of infinite series expansions) for a spherical scatterer, the results are found to be of considerable qualitative relevance for scatterers of other shapes as well. I will briefly relate here how a few important features of the scattered radiation undergo a gradual transformation as the size of the scatterer is made to increase gradually.

For a sufficiently large size of the scatterer, the scattered waves originating from the different parts belonging to it and emitted in any given direction, possess a degree of mutual coherence, and their superposition is responsible for the distinctive features of Mie scattering.

One striking difference from Rayleigh scattering is that, as the scatterer becomes larger in size, the relative preponderance of the smaller wavelengths in the scattered radiation is gradually evened out till, for a size ~ 10 to 100 times the wavelength, *all wavelengths are scattered equally*. This explains the white color of clouds where all the components

of sunlight are scattered equally by the aggregates of water molecules in these clouds.

Another distinctive feature of Mie scattering is a *lack of symmetry between the scattering in the forward and backward directions*, the scattering in the forward direction being relatively more pronounced, which increases with an increase in the size of the scatterer. What is more, for a scatterer of sufficiently large size, the angular distribution of scattered radiation possesses a number of *maxima and minima*, resembling the maxima and minima in the intensity distribution in a *diffraction pattern* (fig. 15-43). Indeed, for a scatterer of size $\sim 10^2$ times the wavelength or larger, the modification of the incident wave by the scatterer can be described as diffraction, where the wave bends around the sphere and, at the same time, fans out to a certain extent away from the forward direction.

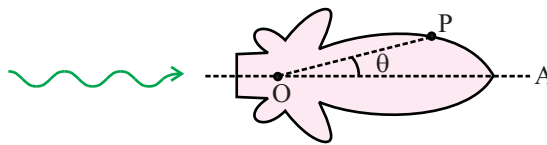


Figure 15-43: Angular distribution of scattered radiation (logarithmic polar plot, with direction of incidence as the polar axis) in Mie scattering with unpolarized incident light (compare fig. 15-41 for Rayleigh scattering, which is the limiting case of Mie scattering for small size of the scatterer); notable features of scattering are the dominance of forward over backward scattering and the maxima and minima in angular distribution); the direction of incidence is OA (compare with fig. 15-41 where this direction is named OC).

15.8.3 Raman scattering

While Rayleigh scattering and Mie scattering (with Rayleigh scattering as one of the limiting cases of Mie scattering) can be looked upon as elastic scattering of photons (the quanta of electromagnetic field, see chapter 16; see also remarks in sec. 15.8.1.1), there may also occur *inelastic* scattering of photons from scatterers (commonly, the molecules of a material), one phenomenon of exceptional importance belonging to this category being *Raman scattering*. This is a process where a photon interacts with a molecule, causing a transition in the *internal state* of the latter, with an attendant change in the energy associated with the internal state, and a corresponding *change in the frequency*

of the photon, as required by principle of conservation of energy. By measuring the frequency shift of the scattered photons, one can learn a great deal about the internal states of the scattering molecules.

A few of the features of the Raman effect can be accounted for by a classical theory where the electromagnetic field is assumed to cause a molecule to develop an oscillating dipole moment, but one whose amplitude itself varies sinusoidally as the molecule undergoes its own vibrational motion. A simple calculation then shows that the oscillating dipole radiates electromagnetic waves with characteristic frequency shifts as observed experimentally.

However, such a classical theory turns out to be an incomplete one, and a more complete semi-quantum (or semi-classical) theory can be formulated where the electromagnetic field is treated classically while the molecule is described in quantum terms. Not much, however, is gained by invoking the quantum theory of the electromagnetic field (though it is convenient to describe the basic process by making use of the concept of photons). Among other things, the semi-quantum theory of Raman scattering explains the occurrence, in the scattered field, of frequencies both higher and lower than the frequency of the incident wave. The components with higher frequencies give rise to characteristic spectral lines observed in an analysis of the scattered wave, referred to as the *anti-Stokes* lines while those with lower frequencies give rise to the *Stokes* lines. In addition, a component with the frequency of the incident field remains in the scattered radiation and corresponds to Rayleigh scattering.

15.9 Wave optics and ray optics

On the face of it, wave optics and ray optics look so unlike each other. Wave optics is all about light propagating in the form of electromagnetic waves, where the oscillations of the associated electric and magnetic field vectors are distributed *throughout* a region of space, the latter often extending to infinite distances. Ray optics, on the other hand, describes light as following definite *ray paths* in space. Yet, for all this seeming difference between the two descriptions, *both* are theories meant for explaining the nature of light.

Where, then, does the reconciliation lie?

The answer has to be sought in the fact that the wavelength of visible light can be considered, in a certain definite sense, to be *small* in a large number of commonly occurring situations, and it is this smallness of the wavelength that allows one to make certain *approximations* that results in the emergence of the ray-theoretic description from the wave-theoretic one.

The seed of this approximation lies in a number of basic characteristics of the *plane wave* solution to the fundamental equations of electromagnetic theory, namely, the Maxwell equations. The wave fronts for a plane wave make up a family of parallel plane surfaces while the wave normals are a set of parallel straight lines perpendicular to these surfaces. As we saw in section 14.4.10, the propagation of electromagnetic energy occurs along these wave normals which can thus be identified as the ray paths.

What is more, a plane wave incident on a plane interface separating two media gives rise to reflected and refracted waves, where the wave normals to the reflected and refracted wave fronts are found to conform to the *laws of reflection and refraction* familiar in ray optics.

For the plane wave, then, the wave normals are indeed the ray paths in an alternative, ray-theoretic description, and in this limited context the ray description is simply a different but equivalent way of looking at the propagation of an electromagnetic wave.

However, the plane wave is an *exceptional* solution - the simplest one - of the equations of electromagnetic theory. This raises the question as to whether the equivalence between the wave theoretic and ray theoretic descriptions that holds in the context of plane waves, can have any relevance in the general context of optics.

This it can, indeed, have. The smallness of the wavelength of light implies that in a certain domain of observations covering a considerable range of optical phenomena, a wave description in terms of what can be *locally* described as plane waves, with small patches of plane wave fronts and correspondingly small segments of wave normals,

enjoys a certain measure of validity.

Fig. 15-44 gives an idea of how the propagation of light can be described in terms of such patches of plane wave fronts. Consider, for instance, a small region around the point P. In this small region the propagation of the wave can be described, under suitable conditions, in terms of a plane wave, and the figure shows a small patch of a plane passing through P which serves effectively as the wave front so far as a description of the wave in the small region under consideration is concerned. The propagation of electromagnetic energy takes place along the path PP' which is a segment of a straight line normal to this patch, where P' is a point located close to P. The electric and magnetic field vectors at P lie on the planar patch and, in conjunction with the unit vector along the local wave normal, form a right handed orthogonal triad.

At P' one again has a similar patch of a plane wave front, with its associated wave normal segment $P'P''$, and so on. These planar patches serving locally as plane wave fronts, and the short segments serving locally as paths of energy flow, join up into smooth surfaces and paths in space, the latter being precisely the ray paths of geometrical optics.

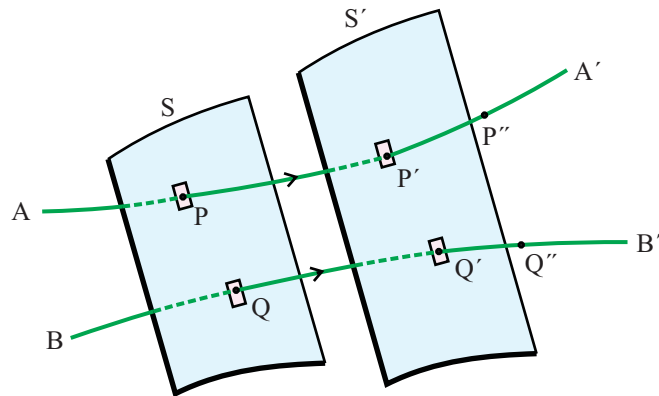


Figure 15-44: Patches showing locally plane wave fronts and corresponding segments of wave normals; these patches and segments join up to smooth surfaces such as S, S' and paths such as AA' , BB' , the latter being the ray paths of geometrical optics.

Such a simplified description in terms of locally plane wave fronts is found to be valid, in a certain approximate sense, in regions where the optical properties of the medium

through which the wave propagates varies sufficiently slowly and which, moreover, are far removed from the sources producing the wave under consideration.

More precisely, the distance over which the optical properties of the medium, such as the refractive index, change appreciably is to be large compared to the wavelength; and the distance of the region under consideration from the sources is to be similarly large.

The surfaces made up of the small patches of locally plane wave fronts are referred to as *eikonal surfaces* or *geometrical wave fronts*. The ray paths are everywhere normal to these surfaces. One can set up mathematical formulas that lead to a determination of these surfaces as also of the ray paths. *These ray paths are then found to conform to all the rules of geometrical optics* stated in section 10.2. In this geometrical optics approximation of wave optics, the transport of electromagnetic field energy takes place along the ray paths with a velocity equal to the local phase velocity v , provided the medium under consideration is a non-dispersive one. For a dispersive medium, the velocity of energy transport, on the other hand, equals the local *group velocity*.

A concise statement of the rules determining the ray paths is referred to as *Fermat's principle*. It includes the principle of rectilinear propagation of light of geometrical optics (which holds for a homogeneous medium), the laws of reflection and refraction, as also the rule for the determination of ray path in an inhomogeneous medium where the optical properties of the medium vary slowly from point to point in it.

The mathematical approximation scheme leading to geometrical optics from the equations of electromagnetic theory also yields a set of rules relating to the variation of intensity along a ray path (intensity rule), and also to the change of the directions of the electric and magnetic field vectors (polarization rules) along any ray path. In other words, *geometrical optics is a complete package that can be said to derive from the principles of electromagnetic theory in the limit of small wavelengths*.

The rules of geometrical optics are relevant even outside the range of wavelengths cor-

responding to visible light. These constitute a convenient approximation scheme for describing the behavior of electromagnetic waves whose wavelengths are small compared to the dimensions of objects (or of the inhomogeneities) they encounter, and in regions far from their source. Examples of the applicability of these rules are to be found in such areas as radar detection and the propagation of radio waves through atmospheric layers.

Chapter 16

Quantum theory

16.1 Introduction

Quantum theory is the theory describing the behavior of *microscopic* objects like molecules and atoms, their elementary constituents, and to some extent, of clusters and aggregates made up of these microscopic objects. By contrast, larger objects like the ones we usually perceive around us in our everyday experience, or even larger ones like the heavenly bodies, are described by *classical* concepts. Newton's laws constitute a basic component of these classical concepts. Thus, a great deal relating to the interaction and motion of these *macroscopic* bodies can be explained quite accurately with Newtonian concepts like force, momentum, and acceleration, together with Newton's laws of motion.

Electromagnetic theory is another basic component of what is known as the classical view of nature, providing us with the explanation of electrical and magnetic phenomena and phenomena relating to the propagation of electromagnetic waves like light, radio waves, and microwaves.

All these classical concepts, while immensely successful in explaining a large body of experience and observations are, however, found to be ineffective in providing consistent answers to a number of fundamental questions relating to the *frontiers* of our everyday

experience, or frontiers encountered in specially designed experiments.

Experiments can be designed that can probe into realms of nature far removed from our sensual world. In this context, one has to distinguish between two different directions in which ingeniously devised apparatus and equipment can extend the horizons of our experience. One of these involves ever larger spatial dimensions and time intervals characteristic of the cosmic world or, what can be termed the *large scale* universe. Explanation of phenomena relating to the large scale universe also need a number of basic modifications of the classical Newtonian ideas but those are of a different nature compared to the radically new ideas required in the other realm, the *microscopic world*. The latter is brought to our senses by experimental observations involving finely tuned instruments. And it is in this microscopic realm that quantum theory gives us a consistent world-view.

This, however, does not mean that quantum theory has no relevance in our world of everyday experience or in the large scale universe, or that the ideas relevant in explaining commonly observed phenomena involving macroscopic objects are not meaningful in the microscopic world. In reality, sets of ideas and concepts applicable to the various realms of experience and observation are not unrelated to one another. Each set of concepts has to be *consistent* with other ones in that there exist certain *overlapping* regions of experience and observation where results derived from more than one conceptual frameworks are to have a certain correspondence with one another.

16.1.1 Quantum and classical concepts: analogy from optics

This can be understood with the help of an analogy drawn from *optics*. Certain phenomena in optics can be satisfactorily explained with the help of concepts relating to *rays*, while certain other observations need the *wave theory* for a proper explanation. Yet, ray optics is not totally disjoint from wave optics, being related to the latter in a very definite manner: ray optics is, in a sense, an *approximation* that can be derived from wave optics, applicable in a certain domain of observation (see section 15.9).

Thus, while the formation of shadows and images can be explained with the help of ray

optics, careful observations show that there occur characteristic variations of intensity around the borders of shadows and images that is satisfactorily explained only with the help of ideas of wave optics. The latter are of a more general validity, applicable in domains of observation where the concepts of ray optics fail to produce satisfactory explanations.

Similarly, the framework made of classical concepts is, in a sense, derived from quantum theory by means of a scheme of approximation and yields explanations of phenomena where these approximations are adequate. Quantum concepts, on the other hand, are applicable over a wider domain of observations including phenomena in the microscopic world. In a manner of speaking, quantum theory constitutes, in the present state of our knowledge, the ultimate physical theory of nature, in the light of which our familiar classical concepts are to be interpreted and evaluated.

Quite naturally, the basic concepts constituting the building blocks of quantum theory appear completely strange and at odds with our familiar classical concepts. This raises a problem for many who feel deeply puzzled while trying to understand these in terms of classical ideas. A better approach is to proceed *the other way round*, interpreting classical concepts in terms of quantum ideas. Of course, there still remains the question of *how*, in the first place, the concepts of quantum theory emerged from the midst of classical ideas.

16.1.2 Emergence of quantum concepts

Though this raises an important issue, it is, nevertheless, an issue in the *history of science*. It is indeed an interesting and important question as to how the radically new concepts of quantum theory could emerge from the womb of the classical theory. Scientists in the early stages of development of quantum theory, while trying to explain certain crucial observations that could not be understood on the basis of classical ideas, did necessarily have to think *in classical terms* while trying to develop new ways of looking at things. They had to do this because a new theory does not emerge as a complete package all in a single day, but is assembled in bits and pieces, on old grounds.

Thus, in the early days, quantum theory used a *mixed* language where a number of new conceptual elements were brought in, retaining, to a certain extent the classical way of looking at things.

This gave rise to grave problems of *interpretation* since the theory at that stage contained ideas of disparate origin. Still, the pioneers in the field persisted with their efforts at building up a new theory, and the *old quantum theory*, made up of these mixed ideas, is now replaced with a new theory that to a large extent, *supersedes* the classical theory. The problems of interpretation still persist, but as I have mentioned above, this is only because one continues to try to understand quantum theory in terms of the more familiar classical ideas. A more prudent course is to get to know the quantum concepts by themselves, with as little reference to common sense classical concepts as possible and then, to try to understand how the rules of classical theory emerge from those of quantum theory as approximations, valid in the realm of observations relating to macroscopic bodies.

Unfortunately, the basic ideas of quantum theory have a considerable degree of mathematical content with which one may not necessarily feel comfortable. I will therefore confine myself to pointing out certain key observations that led people to search for a new theory and to outlining the new concepts they introduced at the early stages of quantum theory, without trying to explain the *logic* of these ideas. I will also state in general terms a few conclusions resulting from the new concepts without trying to substantiate these concretely in a deductive manner. I hope to thereby give you an idea as to where quantum rules contrast with classical ones, and to launch you onto a more complete study of quantum theory.

16.2 Quantum and classical descriptions of the state of a system

The first major difference between classical and quantum systems I want to impress upon you relates to the concept of the *state* of a system, and *values of observable quan-*

tities in a given state. For this imagine an idealized system, namely a single particle moving along a straight line. In classical theory the state of the particle at any given instant of time is completely specified by its position co-ordinate (x) and its velocity, or its momentum (p), while other systems may need more than one position co-ordinates and momentum components for the specification of the instantaneous state. Any function of x and p constitutes an observable quantity relating to this simple idealized system, and possesses a specific, well defined value in a state of the system. For instance, if $V(x)$ be the potential energy of the particle at position x , then its total energy is an observable quantity represented by the function

$$E = \frac{p^2}{2m} + V(x), \quad (16-1)$$

where m stands for the mass of the particle. If one measures the values of the position co-ordinate and the momentum in any given state then one will obtain specific values x and p , and a measurement of energy will also yield a correspondingly specific value.

In quantum theory, however, it is not possible to describe the state of the particle at any given instant of time in such simple terms, i.e., by specifying the values of x and p . In the following (sections 16.2.3 and 16.2.4) I will briefly outline the basic idea as to how the state is described in quantum theory, and will only point out here that a state does *not* necessarily correspond to well defined, specific values of x and p . Instead, if one measures, say, x , in a given state of the system, and repeats the measurement a large number of times, taking care that every time the particle is in the *same* state, one will find that the measured values correspond to a *random variable*.

The idea of a random variable may be familiar to you from what you learned in the kinetic theory of gases (see section 8.24), and is conveniently introduced by the example of the throw of a die. As you throw a die, you cannot predict with certainty what number is going to come up. Instead, any number, from 1 to 6, is a possibility. However, for any given die, you *can* form an idea of the *relative frequencies* with which the numbers turn up in a *large number* of throws of the die. For a straight (or unloaded) die thrown N times, for instance, any one of the six numbers will appear close to $\frac{N}{6}$

times, provided N is large enough. One expresses this by saying that each number on the die has a *probability* $\frac{1}{6}$ associated with it.

More generally, a random variable has a number of possible values associated with it (the numbers 1 through 6 in the above example) and a probability associated with each of these possible values ($\frac{1}{6}$ for each of the six values in the above example). The latter is said to constitute the *probability distribution* of the random variable in question. The sum of all the probabilities involved in the probability distribution has to be unity (reason out why).

Similarly, the measured values of momentum or of any other observable quantity like, for instance, the energy, will also correspond to a random variable. Each observable is characterized by its own set of possible values and probabilities associated with these values, the latter depending on the state under consideration. There may exist special states for which some particular observable has a precise, well defined value, but then other observable quantities are again characterized by their respective measured values coming up like values of random variables.

16.2.1 Illustration: the free particle in one dimension

As an interesting illustration of these features of a state of a quantum system and of the measured values of observable quantities in a state, let us focus on a *free particle*, i.e., a particle for which $V(x) = 0$ (any constant value of $V(x)$, not necessarily zero, would also correspond to a free particle because the force on the particle would be zero). Imagine a state of the particle with a well defined value (p) of the momentum. As mentioned above, such special states of the system are possible, the state mentioned above being one in which a measurement of the momentum yields with certainty the value p . A special feature of a free particle is that, in such a state, the *energy* of the particle *also* possesses a definite value, namely $E = \frac{p^2}{2m}$. In other words, if one measures the value of the energy a large number of times, every time making sure that the particle is in the same state as the one mentioned above, one will invariably end up with the value $\frac{p^2}{2m}$. This much sounds like features I mentioned above for the classical description of the system. But

the resemblance ends here. For, supposing now that one measures the *position* of the particle in the state under consideration, one will face a startling situation : repetitions of the measurement *do not all yield the same value*. As a matter of fact, a large number of repetitions will be found to yield *all possible values* ranging from $-\infty$ to ∞ , and with uniform probability!

Admittedly, this is too idealized a system and too special a state that we are considering (and the ‘measurements’ referred to above are also idealized ones), but nevertheless, it does illustrate strikingly how the quantum description contrasts with the classical description of systems. Our common sense notions tell us that the particle *must* be located *somewhere* at any given instant of time, and *that* location has to be a specific one. How then, can a measurement of its position yield all possible values from $-\infty$ to ∞ ? But it is precisely this counter-intuitive fact that one will *have to* come to terms with if one aims at describing consistently all the observed phenomena in the microscopic world. And all attempts to explain this and other facts relating to the microscopic world *in terms of* familiar concepts of the classical theory are bound to fail. This is a consequence of an *innate richness* of the concepts necessary to describe adequately the phenomena of the microscopic world as compared to those relating to the commonly experienced macroscopic world.

Again, an analogy from optics may help understand the situation. Imagine a ray, selected from a bundle of rays, by means of a pin-hole, and a photographic plate placed in the path of a ray. One expects a single spot on the plate where the ray hits it or, at most a small extended spot since the pin-hole, having a small but finite size, lets pass a narrow bundle of rays. In reality, the ray path through any given point is nothing more than an approximate indicator of the direction of propagation of energy past that point, and it is a wave, extended in space, that hits the photographic plate, resulting in the formation of a *diffraction pattern* on the photographic plate. One can understand the ray as an approximate description of wave features but not the other way round - there is no way one can understand the wave by invoking the simpler features of the ray.

16.2.2 Wave-like features

A free particle with a definite value of the momentum (p) is then equally likely, at any given instant of time, to be found anywhere on the x-axis (the straight line along which it has been assumed to be moving) and it is in this feature that the particle resembles a *wave* because a wave, say, on the surface of water, is not usually confined to one single small region but is found to be spread throughout the surface. In other words, the behavior of a particle in the quantum world is characterized by subtle and rich features - the closest way we can describe these in terms of familiar notions is to say that its behavior combines a number of features known of a particle with those of a wave. Indeed, the analogy with a wave goes deeper - one can characterize the state of the particle (the free particle with a well defined value of momentum in the present instance) in terms of a plane monochromatic wave with a *wavelength* (λ) and a *frequency* (ν).

Recall that the plane monochromatic wave is a basic, if idealized, form of wave motion encountered in numerous areas in physics such as acoustics and optics. For instance, the excess pressure for a plane monochromatic wave propagating along the x-axis varies with position and time as

$$p = p_0 \exp(i(kx - \omega t)), \quad (16-2)$$

where I have employed a complex representation of the wave, the actual value of the excess pressure being the real part of the above complex-valued expression. In this representation, the wave number k and the angular frequency ω are related to the wavelength and frequency as

$$k = \frac{2\pi}{\lambda}, \quad \omega = 2\pi\nu. \quad (16-3)$$

The reason I mention this here is that the special state of a free particle moving in one dimension with a specific, well defined value (p) of the momentum is described in quantum theory with a mathematical expression precisely similar to (16-2) though the expression now does not stand for excess pressure but relates to what is known as the

wave function in quantum theory and, moreover, it is the complex expression, and not just its real part that is of physical relevance in the quantum context. As I mentioned earlier, I will not go into detailed considerations relating to wave functions and their physical relevance, but will only state that it is this wave function that determines the *probabilities* for various values obtained in measurements of observable quantities in any given state of the system. And it is in this sense that the wave function completely describes the state of the system.

16.2.3 Wave function: de Broglie relations

With this in mind, I now rewrite (16-2) so as to represent the wave function of the free particle with a definite value of the momentum:

$$\psi(x, t) = A \exp(i(kx - \omega t)), \quad (16-4)$$

where ψ stands for the wave function and A is a constant.

The question that now comes up is, what would the values of k and ω be so that (16-4) can correctly describe the state of the free particle under consideration? The answer to this question is provided by a famous principle propounded in early days of quantum theory by *Louis de Broglie* in the form of the following relations:

$$k = \frac{p}{\hbar}, \quad \omega = \frac{E}{\hbar}, \quad (16-5)$$

where \hbar is a constant related to *Planck's constant* ($h = 6.626 \times 10^{-34}$ J·s) as

$$\hbar = \frac{h}{2\pi}. \quad (16-6)$$

16.2.4 Quantum description of state: summary

In summary, the description of the state of a system in quantum theory differs radically from that in classical theory and requires the knowledge of the appropriate *wave function*. This description involves subtle and rich aspects that cannot be interpreted

in classical terms. The value of any observable quantity in a given state obtained by measurements turns out, in general, to be a *random variable*, with a certain probability distribution over a set of possible values. The wave function determines all the probabilities relating to measurements of observable quantities in the state under consideration, providing the most complete information possible about the system in the given state.

One way to describe this behavioral aspect of a quantum system is to say that it combines features of a particle (or a system of particles) with those of a wave - the two sets of features being contrary to each other from the classical point of view. A special and idealized state of a simple quantum system is that of a free particle moving along a straight line and having a well defined value of the momentum, wherein the energy also happens to have a well defined value. However, in this state, a measurement of the position shows that the particle, at any given instant of time, can be found *anywhere* on the line with uniform probability - a feature reminiscent of a plane monochromatic wave. This corresponds to a wave function describing the state given by (16-4). The energy and momentum values are related to the wavelength and frequency of the wave by the de Broglie relations (16-5), together with (16-3).

A free particle in one single dimension can be found in *other* states as well, and any such state can again be described in terms of a wave function, though the latter may in general differ from the plane monochromatic wave (16-4). These ideas can be generalized to describe states of systems involving more than one particles, where the particles move in all three dimensions of space. In all these instances, the relevant wave functions determine the probabilities of various possible values coming up when measurements are made for an observable quantity relating to the system under consideration. These probabilities make up all that quantum theory can tell you about the state of the system in the microscopic world.

It is important to mention here that the reason why the term 'wave function' is employed is that the latter appears as the solution of a certain differential equation which, from a mathematical point of view, is analogous to the *wave equations* encountered in other areas of physics such as those relating to acoustic waves (chapter 9) and electromagnetic

waves (chapter 14). This differential equation, describing the time-evolution of the state of a quantum system, is referred to as *Schrödinger's equation* (refer to sec. 16.6 below).

16.3 The principle of uncertainty

Having said all this, I have to make it clear that the state of a free particle described by a wave function resembling a plane monochromatic wave is an idealization since in reality it is not possible to prepare a particle in a state with a *precise* value of the momentum. Imagining a charged particle with a charge q , for instance, the set-up designed to impart a given momentum to the particle is to accelerate it through a certain potential difference, say V , such that it acquires an energy

$$E = qV, \tag{16-7a}$$

and correspondingly a momentum

$$p = \sqrt{2mE} = \sqrt{2mqV}. \tag{16-7b}$$

where m stands for the mass of the particle.

16.3.1 Uncertainty in momentum

However, in this set-up there remain a number of sources for inaccuracy whereby the momentum acquired by the particle may differ from the value given by formula (16-7b) like, for instance, the initial momentum of the particle before it starts being accelerated by the voltage. While some of these sources operate in an uncorrelated way, or *incoherently*, some others may operate *coherently*, i.e., in a correlated manner. In the latter case, these sources of uncertainty give rise to a state of the particle similar to one described by a plane wave but with some difference. The difference relates to the shape of the wave profile and a consequent *uncertainty* in the momentum of the particle.

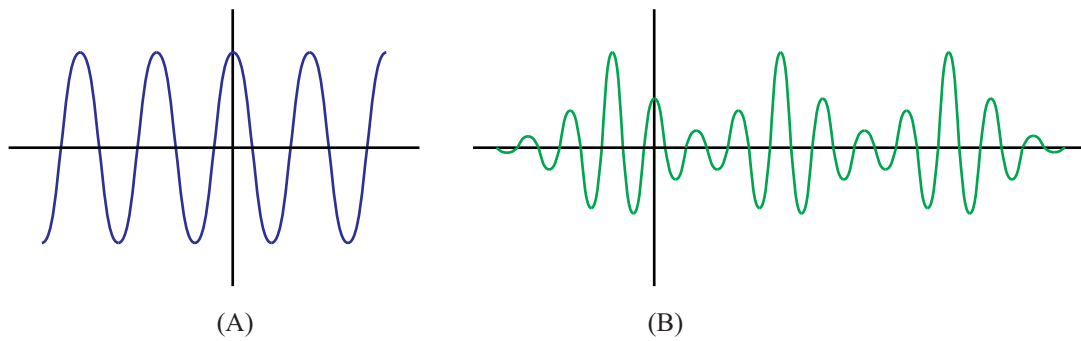


Figure 16-1: Schematic representation of the real part of the wave function of a free particle in a state with (A) well defined momentum and energy, and (B) a spread in possible values of momentum; while the former is a regular sinusoidal curve, the latter may look irregular; each curve plots the wave function against the position co-ordinate x at any specified instant of time.

In other words, the resulting state is one where a measurement of momentum now does not yield one single precise value, but various possible values, with a certain probability distribution. This is why I said earlier that a state with a precise value of the momentum is an idealized one.

Figure 16-1(A) shows schematically the wave function (only the real part of the complex function is shown for the sake of presentation) for a free particle in a state characterized by a well defined momentum and energy while fig. 16-1(B) depicts, again schematically, the wave function in a different possible state where the momentum does not have a precise value. While the former is a regular sinusoidal curve, the latter, though wavy, does not have a sinusoidal waveform. Thus, in (A) the oscillations continue for $x \rightarrow \pm\infty$ while in (B), the oscillations die out at large distances. A wave function of the form of fig. 16-1(B) is sometimes referred to as a *wave packet*.

Depending on the probability distribution over the various possible momentum values, one can calculate quantitatively the *spread* of momentum, or momentum *uncertainty* in the state under consideration (refer to sec. 16.4 for a brief introduction to the relevant concepts). Calling this spread Δp , one can now see how this spread affects the result of measurement of the *position co-ordinate* of the particle. And here lies one big difference distinguishing the rules of the quantum world from those of the classical world.

16.3.2 Momentum and position uncertainties

In the classical theory, the result of measurement of one observable quantity like, say, the momentum does not have any effect on the result of measurement of an independent observable quantity like the position. However, in quantum theory, the spread in momentum values for a given state *does turn out to have a relation* with the spread in the position co-ordinate of the particle in the same state. In the plane wave state (16-4) the spread in momentum is *zero* since the momentum has a precisely defined value. On the other hand, the spread in position (Δx) is *infinite* since in that state the particle happens to be equally likely to be found anywhere on the x-axis. As the momentum spread now increases to Δp , the spread in position is found to *decrease* to some other value, i.e., the particle is now more likely to be found in a region of space of lesser extent. This *inverse* relation between Δp and Δx is made more precise by a relation of the form (refer, once again, to sec. 16.4 for relevant definitions)

$$\Delta p \Delta x \geq \frac{\hbar}{2}, \quad (16-8)$$

which tells us that in any state whatsoever of the particle (the statement can be generalized to other systems obeying quantum principles) the *product* of the uncertainties in position and momentum has to be larger than a certain minimum value, determined by the Planck constant.

The relation (16-8) is the mathematical expression of the *principle of uncertainty* in quantum theory. For a particle moving in three dimensions the principle involves three relations of the above form, one for each component of position and momentum. The terms ‘uncertainty’ and ‘spread’ are used synonymously to denote the extent to which the measured values of a quantity deviate from its mean value. For instance, if the probabilities of large deviations from the mean value are small, then the uncertainty is also small, while if the probabilities of such deviations are relatively large, then the uncertainty is also correspondingly large.

Problem 16-1

A proton of mass $m = 1.66 \times 10^{-27}$ kg and of kinetic energy $E = 1.0$ keV moving along the x-axis of a Cartesian co-ordinate system, is made to pass through a long slit in the y-z plane, parallel to the z-axis, the width of the slit parallel to the y-axis being $a = 3.0 \times 10^{-5}$ m. If Δp_y be the uncertainty in the y-component of the momentum of the proton consequent to passing through the slit, and p_0 denotes its momentum along the x-direction, estimate the value of $\frac{\Delta p_y}{p_0}$.

Answer to Problem 16-1

HINT: A proton with energy less than a value of the order of an MeV can be described in non-relativistic terms. Thus, the momentum p_0 of the proton along the x-axis is given by $p_0 = \sqrt{2mE}$. As the proton passes through the slit, the uncertainty Δy in its position along the y-direction can be assumed to be of the order of a , the width of the slit. According to the principle of uncertainty, the uncertainty Δp_y is then of the order of $\frac{h}{2\Delta y} = \frac{h}{4\pi a}$. In other words, an estimate for the ratio $\frac{\Delta p_y}{p_0}$ would be $\frac{h}{4\sqrt{2\pi a}\sqrt{(mE)}}$. Making use of the known value of h (6.626×10^{-34}) J·s, and of the given values of m , a , E ($= 10^3 \times 1.6 \times 10^{-19}$ J), the required estimate is seen to be of the order of 2.4×10^{-9} (approx).

NOTE: (1) A non-relativistic description of the dynamics of a particle is valid when its velocity is small compared to c , the velocity of light in free space. For a proton of energy 1.0 keV ($1\text{eV} = 1.6 \times 10^{-19}$ J), the velocity works out to 4.4×10^5 m·s⁻¹ (check this out). (2) The assumption that the uncertainty Δy is of the order of a , the width of the slit is in the nature of a rough estimate. (3) In estimating Δp_y , we have used the equality sign in eq. (16-8); this sets the lower limit in the estimate of $\frac{\Delta p_y}{p_0}$.

The next section (sec. 16.4) is aimed at giving you a number of concrete and precise statements regarding the probability distribution of values of an observable quantity and the uncertainty in the measurement of that quantity in any given quantum state of a system.

16.4 Observable quantities, probability distributions, and uncertainties

We continue with the simple quantum system corresponding to a particle whose motion is restricted to the x-axis of a co-ordinate system. Generalizations to systems involving

more than one particles, all moving in three dimensions are straightforward, but will not be considered here separately.

As I have already mentioned, a *state* of the system is described by means of a wave function ψ , which is in general complex-valued, its value at the point x being $\psi(x)$. Thus, you have to distinguish between the wave function and its value at any point x . The former is an abstract entity that can be expressed in terms of a variable other than x . More specifically, the mathematical characterization of the wave function describes it as a *vector*. The notion of a vector as an element in a linear vector space was mentioned briefly in a note in section 2.1, where it was mentioned that each vector space is characterized by a certain *dimension*. The vector space made up of the wave functions (describing all possible states) of a quantum system is special in that it is *infinite dimensional*.

Turning our attention now to the *observable quantities* (such as position, momentum, and energy) of the system, these also belong to a special mathematical category termed *linear operators*. More specifically, each such observable corresponds to a linear operator in the vector space made up of the wave functions, where a linear operator acts on a vector to produce another vector in the same vector space. The operator corresponding to any observable is usually denoted by means of a hat placed over the symbol for that observable. Thus, the operator corresponding to the position co-ordinate, which is an observable, is denoted by \hat{x} , while the operators for momentum and energy are denoted by \hat{p} and \hat{H} where H stands for the energy observable (commonly referred to as the *Hamiltonian* of the system under consideration) given by the expression (16-1) in the classical description. In the following, however, we will omit the hat symbol for the sake of brevity.

As mentioned earlier, given a state ψ (a vector in the mathematical description), the measured values of any observable, to which there corresponds an operator, say, A (recall that we have opted not to use the hat symbol as an explicit reminder of the fact that this is, in the mathematical description, a linear operator) correspond to a random variable, with possible values, say, a_1, a_2, \dots (all these being real numbers) with a certain probability distribution, with the probability, say, P_i for the value a_i ($i = 1, 2, \dots$). What

is important to mention here is that the possible values a_i depend only on the observable A , and not on the state ψ , while the probabilities P_i do depend on ψ . Incidentally, the probabilities P_i ($i = 1, 2, \dots$) characterizing the probability distribution for any observable quantity A and any state ψ are to satisfy the requirement

$$\sum_i P_i = 1, \quad (16-9)$$

which is referred to as the *normalization* condition for the probabilities (refer to formula (8-117b); a number of basic considerations relating to random variables and probabilities have been outlined in section 8.24).

Though the possible values of the observable A have been indicated above to form a *discrete* set of numbers (a_1, a_2, \dots), these may, more generally, be *continuously* distributed as well. For instance, the possible values of measurement of the position observable x (hat omitted!) range continuously from $-\infty$ to $+\infty$, while the momentum observable also has its possible values so distributed. For any specified observable A , the mathematical rules of quantum theory tell us what the possible values a_1, a_2, \dots will be, while the probabilities P_i can also be known from the working rules of quantum theory once the state ψ is specified.

For instance, the probability of the measured value of the position observable lying within x and $x + \delta x$ (note that *this* x is not to be denoted with a hat on it since it is just a real number - a possible value in the measurement of the observable x (properly to be denoted by \hat{x}); we assume that δx is a sufficiently small interval in which the *probability density* $P(x)$ introduced below does not vary appreciably) is given by

$$P(x)\delta x = |\psi(x)|^2\delta x. \quad (16-10)$$

The fact that the possible values of measurement of the position observable are distributed continuously requires that the probability distribution of the measured values be described in terms of a probability *density* $P(x)$ since now it does not make sense to specify the probability for a sharply defined value. The formula (16-9) now appears as

the requirement

$$\int_{-\infty}^{\infty} |\psi(x)|^2 dx = 1, \quad (16-11)$$

which is referred to as the normalization condition on the wave function ψ .

Given the set of possible values $\{a_i\}$ and the probability distribution $\{P_i\}$, the *mean* value (or *expectation* value) of the observable A in the state ψ can be worked out as

$$\langle A \rangle_\psi = \sum_i a_i P_i, \quad (16-12)$$

where the suffix ψ has been attached to $\langle A \rangle$ in order to indicate that it is the state ψ that the probability distribution $\{P_i\}$ refers to. Often, however, the suffix is omitted in favor of the simpler notation $\langle A \rangle$, without explicit reference to the state ψ .

The expectation value gives the average of the measured values for a large number of measurements of A in the state ψ . Another quantity of relevance is the *variance*, which tells us how the measured values are spread out around the expectation value. Corresponding to the observable A , we consider the *squared* observable A^2 , for which the set of possible values obtained on measurement are a_1^2, a_2^2, \dots , with the *same* set of probabilities P_1, P_2, \dots for measurements in the state ψ as for A itself (for instance, if the probability for the value 4 to turn up in the throw of a die be $\frac{1}{3}$, then the probability for the squared value 16 of the square of the number turning up in a throw will also be $\frac{1}{3}$, because the two probabilities actually refer the *same* event). Thus, the mean value of A^2 in the state ψ will be

$$\langle A^2 \rangle = \sum_i a_i^2 P_i, \quad (16-13)$$

where the suffix ψ has been omitted for the sake of brevity (the more complete expression would be $\langle A^2 \rangle_\psi$). The variance is then defined as

$$var(A) = \langle A^2 \rangle - \langle A \rangle^2, \quad (16-14)$$

which actually stands for the square of the *standard deviation* of the statistical distribution of the measured value of A in the state ψ . It is the standard deviation, or the square root of the variance that is referred to, in quantum theory, as the *uncertainty* in the measurement of A (refer back to sec. 16.3), and is commonly denoted by ΔA (example: position and momentum uncertainties Δx , Δp featuring in the formula (16-8) expressing the *principle of uncertainty* in quantum theory):

$$\Delta A = \sqrt{\langle A^2 \rangle - \langle A \rangle^2}. \quad (16-15)$$

All these formulae find their application to work out the expectation values and uncertainties of the position observable x and the momentum observable p in any given state ψ , which is a function of the real variable x .

For instance, the expectation value of the position observable is given by

$$\langle x \rangle = \int_{-\infty}^{\infty} x |\psi(x)|^2 dx, \quad (16-16)$$

which is obtained by combining (16-10) with (16-12) and by noting that the summation in the latter is to be replaced with an integral extending from $x \rightarrow -\infty$ to $x \rightarrow \infty$. Likewise, the expectation value of the squared position variable x^2 is given by

$$\langle x^2 \rangle = \int_{-\infty}^{\infty} x^2 |\psi(x)|^2 dx. \quad (16-17)$$

The position uncertainty Δx in the state ψ is then obtained by using equations (16-16) and (16-17) in (16-15).

In the case of the momentum observable p , the analogous formulae look a bit different, which I will now state for your reference

$$\langle p \rangle = \int_{-\infty}^{\infty} \psi(x)^* \frac{\hbar}{i} \frac{d}{dx} \psi(x) dx, \quad (16-18a)$$

$$\langle p^2 \rangle = \int_{-\infty}^{\infty} \psi(x)^* \left(\frac{\hbar}{i}\right)^2 \frac{d^2}{dx^2} \psi(x) dx, \quad (16-18b)$$

from which the uncertainty Δp can be worked out by referring, once again, to (16-15).

The uncertainties Δx and Δp , worked out through these formulae, satisfy the inequality (16-8), no matter what the state ψ happens to be. A state for which the uncertainty product $\Delta p \Delta x$ attains its minimum possible value, i.e., $\frac{\hbar}{2}$, is referred to as a *minimum uncertainty state*.

This seems to be a good place to try out a few exercises.

Problem 16-2

The probabilities of numbers 1, 2, ..., 6 turning up at the throw of a die are, respectively $\frac{1}{4}, \frac{1}{6}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{4}$. Work out the mean and standard deviation of the numbers turning up in a large number of throws.

Answer to Problem 16-2

HINT: Let the number turning up in a throw be denoted by A . Then, by formula (16-12), the mean is seen to be $\langle A \rangle = 1 \times \frac{1}{4} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{12} + 5 \times \frac{1}{12} + 6 \times \frac{1}{4} = \frac{10}{3}$. Likewise, the mean value of A^2 works out to $\langle A^2 \rangle = \frac{89}{6}$. The standard deviation is then $\Delta A = \sqrt{\langle A^2 \rangle - \langle A \rangle^2} = \sqrt{\frac{67}{18}} \approx 1.93$.

Problem 16-3

Consider a state of a particle in one dimensional motion, given by the normalized wave function $\psi(x) = \left(\frac{a}{\pi}\right)^{\frac{1}{4}} e^{-\frac{a}{2}x^2}$, where a is a real positive parameter. Making use of the integrals $I \equiv \int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}$, and $J \equiv \int_{-\infty}^{\infty} x^2 e^{-ax^2} dx = \frac{1}{2a} \sqrt{\frac{\pi}{a}}$, obtain the value of $\Delta p \Delta x$ for this state, and compare with (16-8).

Answer to Problem 16-3

HINT: For the given wave function ψ , one finds $\langle x \rangle = \sqrt{\frac{a}{\pi}} \int_{-\infty}^{\infty} x e^{-ax^2} dx = 0$ since the integrand is an odd function of x . Likewise, making use of (16-18a), we obtain $\langle p \rangle = \sqrt{\frac{a}{\pi}} \frac{\hbar}{i} \int_{-\infty}^{\infty} (-ax) e^{-ax^2} dx = 0$. Again, by making use of the integral J , one gets $\langle x^2 \rangle = \sqrt{\frac{a}{\pi}} \int_{-\infty}^{\infty} x^2 e^{-ax^2} dx = \sqrt{\frac{a}{\pi}} J = \frac{1}{2a}$.

while, likewise, $\langle p^2 \rangle = -\hbar^2 \sqrt{\frac{a}{\pi}} (a^2 J - aI) = \frac{\hbar^2}{2} a$. One then finds, $\Delta p \Delta x = \frac{\hbar}{2}$. Comparing with the relation (16-8), one observes that the state ψ , described by a *Gaussian* wave function, i.e., one of the specified form, corresponds to the minimum possible value of the product $\Delta p \Delta x$ consistent with the principles of quantum theory, regardless of the parameter a , which describes the ‘width’ of the Gaussian function. Such a state is referred to as a *minimum uncertainty state*.

Problem 16-4

Imagine a quantum mechanical system for which a measurement of energy yields two possible values $E_1 = 0$ and $E_2 = \epsilon$, with probabilities $\frac{3}{4}$ and $\frac{1}{4}$ for a given state ψ . Calculate the mean energy and the energy uncertainty in the state.

Answer to Problem 16-4

HINT: Denoting the probabilities of energies E_1 and E_2 by $P_1 = \frac{3}{4}$ and $P_2 = \frac{1}{4}$, we find, on invoking formulas (16-12) and (16-13), with the energy for the observable A , $\langle E \rangle = \frac{1}{4}\epsilon$, and $\langle E^2 \rangle = \frac{1}{4}\epsilon^2$. This gives, according to formula 16-15, $\Delta E = \frac{\sqrt{3}}{4}\epsilon$.

16.5 The simple harmonic oscillator

Recall the oscillatory motion of a simple pendulum where, for small amplitudes, the bob of the pendulum makes a to-and-fro movement along a straight line, characterized by a fixed frequency. A similar motion is observed for a small mass attached to one end of a spring, fixed at the other end, where the spring is stretched by a small length and then let go. In each of these instances the motion is characterized by an energy function of the form (16-1) in which the potential energy $V(x)$ is given by

$$V(x) = \frac{1}{2} m \omega^2 x^2, \quad (16-19a)$$

where ω stands for the angular frequency of the oscillator (related to its frequency ν as

$\omega = 2\pi\nu$). The total energy function of the oscillator is thus

$$H = \frac{p^2}{2m} + \frac{1}{2}m\omega^2x^2, \quad (16-19b)$$

where I have used the symbol H because the energy function is often referred to as the *Hamiltonian* of the system under consideration (refer back to sec. 16.4).

16.5.1 Bound system: quantization of energy

One big difference between the oscillator and the free particle is that the motion of the former is *bounded* while that of the latter is *unbounded* - whatever be the amplitude of motion of the oscillator, however far it moves away from the mean position, it always returns to the mean position because of the restoring force pulling it back, so as to repeat its motion. On the other hand, the free particle is under no compulsion to turn back and eventually moves away to arbitrarily large distances from its initial location. This distinction between the two systems gives rise to quite a remarkable consequence: *the quantization of energy* of the harmonic oscillator, as compared with a continuous distribution of the possible values of the energy of a free particle.

While a free particle in the state described by the wave function (16-4) has a well defined, specific value for its energy, that value itself can be anything from zero to infinity. For instance one can have a free particle with an energy $10^{-7}J$ while one with an energy, say, $10^{-6}J$, or any value in between, is equally conceivable. One expresses this by saying that the possible energy values of the free particle are continuously distributed from 0 to ∞ , though in a state described by a wave function of the form (16-4) a measurement of the energy of the particle will yield with certainty only one of these possible values. In a state with a wave function of a different form, a measurement of the energy is more likely to yield a value described by a random variable, but the *likely* values of that random variable will be found to be distributed over a continuous range, *any* value in that range being a possible result of the measurement of energy.

For a harmonic oscillator, on the other hand, the result of measurements of its energy in any given state will once again be described by a random variable, but now the possible

values of this random variable make up a *discrete set*. More specifically, the possible energy values of the oscillator are given by the formula

$$E_n = \hbar\omega\left(n + \frac{1}{2}\right) \quad (n = 0, 1, 2, \dots). \quad (16-20)$$

In other words, the measured energy of the oscillator can be $\frac{1}{2}\hbar\omega$, $\frac{3}{2}\hbar\omega$, $\frac{5}{2}\hbar\omega$, and so on, but it *cannot* be, say $\frac{3}{4}\hbar\omega$, or any value in between, say $\frac{5}{2}\hbar\omega$, and $\frac{7}{2}\hbar\omega$. This is what one means while saying that the energy of the oscillator is quantized. Any energy value belonging to the set (16-20) can be termed an *allowed* energy value, implying thereby that the rules of quantum theory do not allow for any *other* energy value that can be obtained in a single energy measurement in an arbitrarily specified state of the oscillator. Each of these allowed energy values is characterized by a non-negative integer n termed the *quantum number*. Given any specific value of the quantum number, there can be found a special state of the oscillator such that a measurement of the energy of the oscillator in *that* state yields a well defined and precise result, namely the corresponding energy E_n . Such a state of a quantum system for which the result of measurement of the energy yields a well defined and precise value rather than a number of possible values like those of a random variable, is termed a *stationary state*.

In summary, then, when the harmonic oscillator is in one of its stationary states, a measurement of its energy will yield a specific value with certainty, that value being one of the set of allowed energy values (16-20) depending on the stationary state in question. On the other hand, if the oscillator happens to be in a state other than a stationary state, a measurement of its energy will yield a value described by a random variable, the possible values of the random variable being again the set (16-20), where the probabilities of occurrence of these values will be determined by the state under consideration.

16.5.2 Digression: the continuous and the discrete

Notice that the possible results of the measurement of the energy of a system does not depend on the state the system is in, being determined solely by the energy function,

or the Hamiltonian, of the system. What depends on the state is the set of probabilities characterizing the frequencies with which these possible values turn up in a large number of energy measurements, all performed on the same state of the particle. This set of likely values of energy is termed the *energy spectrum* of the system under consideration. What singles out a stationary state is that, for such a state, only one of the values making up the energy spectrum results from an energy measurement, i.e., the probability for that particular value is unity while the probabilities for all other possible values are zero. Referring to what I said about the free particle and the harmonic oscillator, one can then make the statement that *the energy spectrum of a system with unbounded motion is continuous while that for a bound system is discrete*.

Similar statements can be made with reference to other observable quantities of the system. In any given state of the system, the likely values of the observable can make up either a discrete or a continuous set. For instance the possible values of the position or the momentum of the system make up a continuous set, ranging from $-\infty$ to ∞ , while those of any component of the *angular momentum* vector make up a discrete set. One expresses this by saying that, like the energy of a bound system, the angular momentum is *quantized*. It is this quantization, expressed through the quantization of energy, that is of crucial importance in explaining the *emission spectrum of a hydrogen atom*. An early breakthrough in the development of quantum theory came in the form of Bohr's work on the spectrum of the hydrogen atom. This we will have a look at in a following section.

16.5.3 Harmonic oscillator: the uncertainty principle at work

All the stationary states of the harmonic oscillator are bound states. The energies of the stationary states, when arranged in increasing order, form a discrete sequence, given by (16-20). These are commonly referred to as the *energy levels* of the oscillator. Counting the lowest of these as the first level, and the successively higher ones as second, third, and so on, the j th level is characterized by the quantum number $n = j - 1$. Figure 16-2 depicts these energies in a diagram where the potential energy of the oscillator is also shown for the sake of comparison.

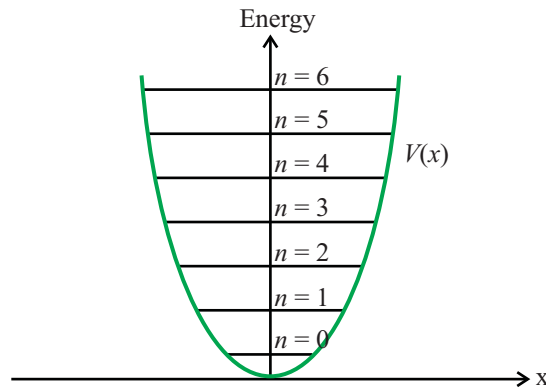


Figure 16-2: The energy levels of the harmonic oscillator, drawn along with the potential energy curve; each energy level is shown as a horizontal line; the successive lines are equi-spaced along the vertical energy axis; the horizontal width of a level indicates the extent of classical motion of the oscillator.

The stationary state corresponding to the lowest energy level (quantum number $n = 0$) is termed the *ground state* of the oscillator, while the states with successively higher energies ($n = 1, 2, \dots$) are termed the first excited state, the second excited state, and so on.

Though the energy of a stationary state such as the ground state is a well defined quantity with a precise value, measured values of other observable quantities are described by random variables, with probability distributions depending on the state under consideration. For instance, if you measure the position of the oscillator a large number of times, every time making sure that the oscillator is in the ground state you will get all results ranging from $-\infty$ to ∞ , with a probability distribution as shown in fig. 16-3. A measurement of the momentum of the oscillator also yields similar results, with a probability distribution looking similar to the graph shown in fig. 16-3.

Making use of these probability distributions one can work out the spreads, or uncertainties, in position and momentum values of the oscillator. One finds that these uncertainties are related to each other as (refer back to formula (16-8))

$$\Delta p \Delta x = \frac{\hbar}{2}. \quad (16-21)$$

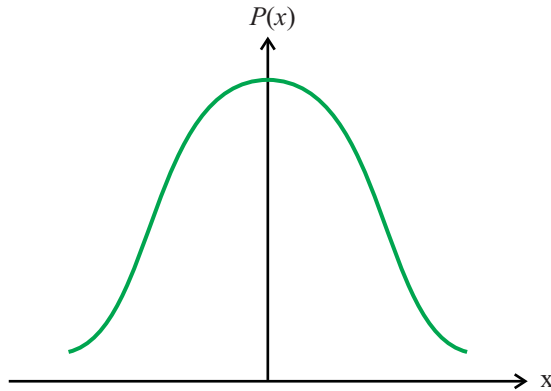


Figure 16-3: Probability distribution (schematic) of various possible values of the position co-ordinate of a harmonic oscillator in the ground state; the ordinate $P(x)$ represents what is termed the probability *density*; for any given position x , and a small interval of length δx , $P(x)\delta x$ gives the probability of the particle being found in the interval x to $x + \delta x$; notice that the particle is more likely to be found close to the origin (the mean position) than further away from it; the probability distribution in momentum looks similar; the graph of the probability density is referred to as a *Gaussian curve*.

Indeed, the ground state of the oscillator is described by a wave function that turns out to be a *Gaussian* one, i.e., of the form of the function $\psi(x)$ given in problem 16-3, where it was found that the state is actually a *minimum uncertainty* one, regardless of the value of the parameter a . The other stationary states also satisfy (16-8) but with higher values of the uncertainty product.

Problem 16-5

A harmonic oscillator of mass $m = 1.66 \times 10^{-27}$ kg has an angular frequency $\omega = 2.5 \times 10^{15} \text{ s}^{-1}$. What is its energy in the stationary state with quantum number $n = 20$? What would be the squared amplitude of oscillation for a classical oscillator (having the same mass and angular frequency) with this energy? Assuming that the energy in the quantum mechanical state is, on the average, equally distributed among the kinetic and potential energies, obtain the mean value of x^2 , the squared co-ordinate, in this state.

Answer to Problem 16-5

The energy in the state with quantum number $n = 20$ is given by the expression $E = \hbar\omega(n + \frac{1}{2}) = \frac{\hbar}{2\pi}\omega(n + \frac{1}{2})$. Making use of the known value of h (6.626×10^{-34} J·s) and given values of ω and n , one gets $E = 5.4 \times 10^{-18}$ J (approx). Since the squared amplitude a^2 of a classical oscillator is related

to its energy as $E = \frac{1}{2}m\omega^2 a^2$, one gets $a^2 = \frac{E}{\frac{1}{2}m\omega^2}$. Making use of given values of m and ω , this is seen to work out to $a^2 = 1.04 \times 10^{-21} \text{ m}^2$.

While the total energy of the oscillator has a specific value in the state with any given quantum number n (20 in the present instance), its kinetic and potential energies are both indeterminate, i.e., are represented by random variables. Each of these, however, has some specific probability distribution associated with it, and hence is characterized by some definite mean or expectation value. Since the averages of the kinetic and the potential energies are the same, both must be half the total energy E . Since, moreover, the expression for the potential energy is $\frac{1}{2}m\omega^2 x^2$, the expectation value of the squared co-ordinate is given by $(x^2)_{\text{av}} = \frac{\frac{E}{2}}{\frac{1}{2}m\omega^2}$. In other words, the expectation value of the squared co-ordinate is *half* the classical squared amplitude, i.e., in the present instance, $5.2 \times 10^{-22} \text{ m}^2$.

16.6 Time evolution of states

In the classical description, the position and momentum of a particle get changed with time in accordance with its equations of motion, which means that the state of the particle evolves in time, depending on its energy function, i.e., the Hamiltonian, since it is the Hamiltonian that determines the equations of motion of the particle.

The way the state of motion of a dynamical system changes in the classical description is commonly expressed in the form of Newton's equations of motion, but these can alternatively be expressed in terms of the Hamiltonian of the system under consideration. For instance, in the simple instance of a particle moving in one dimension, the equation of motion can be broken up into the following two equations, written in terms of the Hamiltonian, which is nothing but the energy function (16-1), and which we now denote by H :

$$\begin{aligned} \frac{dx}{dt} &= \frac{\partial H}{\partial p}, \quad \frac{dp}{dt} = -\frac{\partial H}{\partial x}, \\ H &= \frac{p^2}{2m} + V(x). \end{aligned} \tag{16-22}$$

Taking note of the relation between the force acting on the particle and its potential

energy,

$$F(x) = -\frac{\partial V}{\partial x}, \quad (16-23)$$

one can verify straightaway that the equations (16-22) are equivalent to the Newton's equation of motion for the particle:

$$m \frac{d^2 x}{dt^2} = F(x). \quad (16-24)$$

The equations (16-22) are instances of what are referred to as the *Hamiltonian equations of motion* for a dynamical system.

A corresponding statement holds in the quantum description of a system as well, where its wave function ψ changes with time in a manner determined by the Hamiltonian (the operator corresponding to the energy of the system). Continuing to refer to the simple instance of a particle moving in one dimension, the time evolution is expressed by the equation

$$i\hbar \frac{\partial \psi(x, t)}{\partial t} = -\frac{\hbar^2}{2m} \frac{\partial^2 \psi(x, t)}{\partial x^2} + V(x)\psi(x, t). \quad (16-25)$$

In this equation, $\psi(x, t)$ stands for the value of the wave function ψ at any chosen point x and at any time instant t , and the left hand side of gives the time rate of change of ψ (multiplied with $i\hbar$). The right hand side, on the other hand, involves the second order spatial derivative at the given time t , and represents a function obtained by the action of the Hamiltonian *operator* \hat{H} (note the hat on H !) on ψ (I do not enter into further details here).

The equation (16-25) is referred to as the *Schrödinger equation* in quantum theory.

For any given quantum system, the Schrödinger equation has the general form

$$i\hbar \frac{\partial \psi}{\partial t} = \hat{H}\psi, \quad (16-26)$$

where the left hand side gives the rate of change of the function ψ with time. The

Hamiltonian operator for a single particle moving in one dimension, acting on ψ at time t effects a transformation from $\psi(x, t)$ to the right hand side of (16-25), this transformation being entirely determined by the form of the classical Hamiltonian function (third relation in equations (16-22)).

16.7 Superposed states in quantum theory

The way the state of a quantum system under given conditions is described, and the rules relating any such state to results of observations made with the system in that state, imply a large number of consequences that appear to be just ‘impossible’ from the ‘common sense’ point of view. On closer scrutiny, however, the ‘common sense’ point of view is seen to be the one formed from our experience and observations in the *macroscopic* world, where the rules of classical physics work well. And the apparent ‘impossibility’ of quantum phenomena is just an expression of the fact that the rules of quantum theory, applicable to *microscopic* systems, are just not the ones of classical physics.

The two sets of rules, one for the macroscopic and the other for the microscopic world, are, however, not really irreconcilable. The classical rules are related to the quantum ones in a certain limiting sense, analogous to the way the principles of ray optics are related to those of wave optics.

One crucial feature of the quantum description of a system is the *principle of superposition* that applies to the wave functions describing possible states of the system. Suppose ψ_1 and ψ_2 are the wave functions describing two different states of a system. Suppose further that these states are such that the measurement of a certain observable quantity A gives with certainty the value a_1 for the state ψ_1 and the value a_2 for the state ψ_2 (recall that the principles of quantum theory allow for such states).

The basic quantum rules then decree that a state described by a wave function of the form $\alpha\psi_1 + \beta\psi_2$ is also possible, where α and β are any two complex numbers, satisfying $|\alpha|^2 + |\beta|^2 = 1$. Such a state is referred to as a *superposed* one made up of the states

ψ_1 and ψ_2 . What is interesting to note about such a superposed state is that the measurement of the observable A with the system in this state does not produce a definitive result but instead the result of the measurement can be described in terms of a random variable, where the possible values of the random variable are a_1 and a_2 , with probabilities $|\alpha|^2$, $|\beta|^2$.

It is this fundamental principle of quantum theory, referred to as the principle of superposition, that leads to very many experimental consequences that appear to be bizarre from our ‘common sense’ point of view. However, I will not describe and analyze them here because my aim in this book is more limited: to give you just an initial acquaintance with quantum theory.

16.8 Mixed states: incoherent superposition

The superposed states introduced in sec. 16.7 can be described as *coherent* superpositions, while *incoherent* superpositions are also possible, corresponding to what are commonly referred to as *mixed states* of quantum systems. The distinction between superposed states (this term is commonly used to describe coherent superpositions, a practice we will follow) and mixed states (i.e., incoherent superpositions) of a quantum system is analogous to coherent and incoherent superpositions of electromagnetic waves.

For instance, a superposition of two plane electromagnetic waves, of the form $A_1 e^{i\delta_1} e^{i(kx - \omega t)} + A_2 e^{i\delta_2} e^{i(kx - \omega t)}$ (with A_1 and A_2 real and positive, for the sake of concreteness) is a coherent one if the parameters $A_1, A_2, \delta_1, \delta_2$ have well defined values (i.e., these are *deterministic* variables). If, on the other hand, these parameters are in the nature of random variables, each with a probability distribution over a set of possible values, then the superposition is a mixed one.

Mixed states of a quantum system can be defined in an analogous manner, where one can imagine a number of states to be superposed with no correlations between them. In such a case of incoherent superposition of two states ψ_1, ψ_2 one cannot define un-

ambiguously the superposition coefficients α, β as in sec. 16.7 while, instead, one can define a pair of *weights* or probabilities P_1, P_2 with which the two states are mixed.

More generally, considering a set of *basic* states ψ_1, ψ_2, \dots such that, in the state ψ_n , a certain observable quantity A has a precisely defined value a_n (recall that such states can be defined in quantum theory), a mixed state can be formed out of these basic states with weights P_1, P_2, \dots , where such a mixed state cannot be represented by means of a wave function ψ . Instead, the set of states ψ_1, ψ_2, \dots , *along with* the weights P_1, P_2, \dots , completely define the mixed state under consideration. By contrast, the states that can be represented by means of wave functions (i.e., the ones we have been considering in the preceding sections in this chapter) are referred to as *pure* states.

16.9 Black body radiation: Planck's hypothesis

Any material body maintained at a given temperature, say, T , emits energy in the form of electromagnetic waves into its surroundings and, at the same time, receives electromagnetic radiation from surrounding bodies. The criterion for the temperature to remain constant, i.e., for thermal equilibrium, is that the rates of emission and absorption of energy are to be equal.

Imagine now a large hollow cavity surrounded by an enclosing wall, where the latter is maintained at a given temperature (T). Part of the radiation emitted by the wall remains inside the cavity and moves back and forth in it, where a continuous process of emission and absorption goes on. Eventually, a condition of equilibrium is reached when the radiation inside the cavity attains a state characterized by a number of features depending only on the volume (V) of the cavity and the temperature (T) of the wall. This is referred to as *black body radiation*.

Strictly speaking, the volume V of the cavity is to be assumed to be infinitely large.

This radiation is made up of components of all frequencies (ν) from zero to infinity, and correspondingly all possible wavelengths (λ ; recall that the frequency and wavelength

of monochromatic radiation are related as $\nu\lambda = c$, where c stands for the velocity of light in vacuum). Each such component, moreover, exists in the cavity in the form of a *standing wave* because of the fact that it has to remain confined within the volume V . The radiation inside the cavity possesses an energy since there is an energy cost associated with the setting up of an electromagnetic field. The total electromagnetic energy inside the cavity is made up of contributions from components with all possible frequencies, i.e., the energy required to set up the standing waves in the cavity with these frequencies.

Rather than talking of the energy associated with any particular frequency, it is more appropriate to talk of the *energy density* (i.e., energy per unit volume of the cavity) associated with a small frequency range, say, ν to $\nu + \delta\nu$. From this, one can work out the energy density *per unit frequency interval* at any given frequency, say, ν . Calling this quantity u_ν , the question arises of determining u_ν (commonly referred to as the black body distribution function) as a function of ν and T . This is known as the problem of black body spectrum, which essentially relates to how the energy of the radiation in thermal equilibrium inside a cavity is apportioned among the various frequency components. Knowing u_ν , the energy in the cavity corresponding to radiation in the frequency range ν to $\nu + \delta\nu$ is obtained as $Vu_\nu\delta\nu$.

However, it was found that the *classical theory could not solve this seemingly simple problem*, i.e., in other words, the results obtained from the classical theory could not account for the experimentally observed features of the black body spectrum. Though the classical theory did give good results for certain frequency ranges, it could not explain the variation of u_ν over the entire frequency range. This variation is shown schematically in figure 16-4. Max Planck, in a celebrated *coup* in the history of physics, found out the correct formula for u_ν which reads

$$u_\nu = \frac{8\pi h\nu^3}{c^3} \frac{1}{e^{\frac{h\nu}{k_B T}} - 1}. \quad (16-27)$$

In this expression, h stands for *Planck's constant*, a fundamental constant of nature, with value 6.626×10^{-34} J·s (approx), while k_B stands for Boltzmann's constant, another

fundamental constant one comes across in thermal physics.

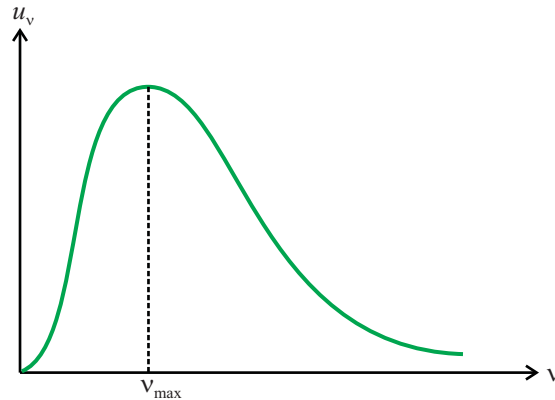


Figure 16-4: Planck distribution (schematic): energy density per unit frequency interval (u_v) as a function of frequency (v); u_v goes to zero for small as well as large values of v ; in between, it reaches a maximum for a certain value of frequency (v_{\max}) that depends monotonically on the temperature.

In arriving at this formula Planck had to make a novel assumption, which was a radical departure from the classical theory. Recall the standing waves of various frequencies making up the black body radiation. Classically, the energy of a standing wave, termed a *mode* in the context of the radiation, can have any value from zero to infinity. However, any component of the electric or magnetic field intensity for a mode at any given point varies sinusoidally, and in this respect a mode resembles a *harmonic oscillator*. This is where quantum theory comes in since, as I have already told you, the quantum theory of the harmonic oscillator differs from the classical theory in that it implies a *discrete* set of possible energy values of the oscillator, any two successive values differing from each other by the amount $h\nu$, where ν stands for the frequency of the oscillator.

The total energy inside the cavity is made up of the energies of the individual modes of radiation. The minimum amount by which the energy of a typical mode of frequency ν can increase is $h\nu$, resulting in an equal increase in the total energy in the cavity. A convenient way to express this increase in energy is to say that an *energy quantum* has appeared in the cavity. Conversely, a decrease in energy by ν can be described as the disappearance of an energy quantum. Such an energy quantum is commonly referred

to as a *photon*. Thus, for instance, if the energy of a mode of frequency ν_1 increases by $2h\nu_1$, and that of another mode of frequency ν_2 decreases by $3h\nu_2$, then one says that two photons of frequency ν_1 have appeared and three photons of frequency ν_2 have disappeared in the cavity.

Planck took into account the *quantum nature of the modes* whereby each mode is essentially a harmonic oscillator, and looked at the black body radiation as a composite system made up of a large number of harmonic oscillators *in thermal equilibrium* at any given temperature T . It is this new approach that resulted in the formula (16-27) describing the black body spectrum.

Along with deriving the formula (16-27), Planck also inaugurated the concept of the photon and of the role of *discreteness* in the behavior of systems. It was indeed an epoch-making concept that was eventually to change the entire face of physics.

Incidentally, Planck's formula is sometimes written in an alternative form, involving the *spectral emittance* of a black body. Let $\delta W(\nu)$ be the rate of radiation of energy per unit area from a black body at temperature T , within a small frequency range ν to $\nu + \delta\nu$ in the normal direction (i.e., at right angles to the emitting surface) per unit solid angle. Then $\frac{\delta W(\nu)}{\delta\nu}$, the rate of radiation per unit area per unit frequency interval at the frequency ν is termed the spectral emittance. The alternative form of Planck's formula expresses $I(\nu)$ as a function of ν :

$$I(\nu) = \frac{2\pi h\nu^3}{c^2} \frac{1}{e^{\frac{h\nu}{k_B T}} - 1}. \quad (16-28)$$

If this expression is integrated over ν over the entire spectrum of electromagnetic radiation, i.e., from $\nu = 0$ to ∞ , and over all possible directions of emission (after multiplying with an appropriate 'obliquity factor'), one obtains the total power radiated by a black body at temperature T per unit area which, according to Stefan's law, is σT^4 , where σ denotes the Stefan constant (refer to sec. 8.23.3.1). One thereby arrives at the following

relation between the constants h and σ :

$$\sigma = \frac{2\pi^5 k_B^4}{15c^2 h^3}. \quad (16-29)$$

16.9.1 Harmonic oscillators in thermal equilibrium

Consider a harmonic oscillator of frequency ν in *thermal equilibrium* with a reservoir at temperature T . A reservoir is a large system characterized by some fixed temperature T that is assumed to remain unchanged when it exchanges energy with a smaller system of interest - the harmonic oscillator in the present instance.

Let the stationary states of the oscillator be ψ_n ($n = 0, 1, 2, \dots$), with energies $E_n = h\nu(n + \frac{1}{2})$. Then the oscillator in thermal equilibrium at temperature T is characterized by a mixed state that can be described as an incoherent superposition of the basic states ψ_1, ψ_2, \dots with weights P_n ($n = 1, 2, \dots$) (refer to sec. 16.8), given by

$$P_n = \frac{1}{Z} e^{-\frac{E_n}{k_B T}}, \quad (16-30a)$$

where

$$Z = \sum_{j=0}^{\infty} e^{-\frac{E_j}{k_B T}}, \quad (16-30b)$$

is referred to as the *partition function* of the oscillator at temperature T ($k_B =$ (Boltzmann constant)). Note that these weights are entirely in accord with *Boltzmann's formula* in statistical physics (refer back to section 8.14, especially to formulae (8-49a), (8-49b)).

One can now describe Black body radiation (sec. 16.9) in the following terms: it is made up of an infinite number of independent harmonic oscillators, corresponding to all possible standing wave modes of electromagnetic radiation in an enclosure, where each oscillator is characterized by some frequency ν (ranging from zero to infinity for the collection of oscillators) and is in thermal equilibrium (with the walls of the enclosure and with all the other oscillators) at some specified temperature T , being characterized by a mixed state of the above description (refer to equations (16-30a), (16-30b)).

It is from this quantum mechanical description of black body radiation that Planck's formula (eq. (16-27)) is obtained.

Problem 16-6

Find the energy in electron volt of a photon for light of wavelength $\lambda = 550 \text{ nm}$.

Answer to Problem 16-6

SOLUTION: A photon can be described in terms analogous to a free particle, where its energy and momentum are related to its frequency and wavelength by the de Broglie formulae, which give, for the energy of the photon, $E = h\nu = \frac{hc}{\lambda}$. Putting $h = 6.626 \times 10^{-34} \text{ J}\cdot\text{s}$, $c = 3 \times 10^8 \text{ m}\cdot\text{s}^{-1}$, and $\lambda = 550 \times 10^{-9} \text{ m}$ (given), one gets the energy E in joule. Since the charge of an electron is $1.6 \times 10^{-19} \text{ C}$, and 1 electron volt is the change in the energy of an electron as it moves through a potential difference of 1 V, we get $1 \text{ eV} = 1.6 \times 10^{-19} \text{ J}$. Thus the required photon energy is $E = \frac{6.626 \times 10^{-34} \times 3 \times 10^8}{550 \times 10^{-9} \times 1.6 \times 10^{-19}} \text{ eV}$, i.e., 2.26 eV (approx).

Problem 16-7

Express Planck's formula (16-27) in terms of wavelength λ in place of the frequency ν .

Answer to Problem 16-7

HINT: If $\delta \tilde{u}(\lambda)$ denotes the energy density of black body radiation at temperature T within the small wavelength range λ to $\lambda + \delta\lambda$, then $\tilde{u}(\lambda) = \frac{\delta \tilde{u}(\lambda)}{\delta\lambda}$ denotes the energy density per unit wavelength range at wavelength λ . If the corresponding frequency range is ν to $\nu + \delta\nu$ then, by definition, $\delta \tilde{u}(\lambda) = \delta u(\nu)$, i.e., $\tilde{u}(\lambda) = u(\nu) \frac{\delta\nu}{\delta\lambda}$. Making use of formula (16-27) for $u(\nu)$, converting from ν to $\lambda = \frac{c}{\nu}$, and ignoring the negative sign in $\frac{\delta\nu}{\delta\lambda}$ (the negative sign arises because of the fact that λ is a decreasing function of ν), one gets

$$\tilde{u}(\lambda) = u(\nu) \frac{c}{\lambda^2} = \frac{8\pi hc}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda k_B T}} - 1}. \quad (16-31)$$

16.10 Bohr's theory of the hydrogen atom

While the harmonic oscillator is a particle moving in one single dimension, an electron in a hydrogen atom moves in *three* dimensions. In the *classical description*, the electron revolves around the nucleus (which can be assumed to be a fixed point due to its relatively large mass) in an elliptic orbit. A special case corresponds to that of a *circular orbit* of radius, say, r , in which the electron moves with a uniform speed, say, v (fig. 16-5). It is the electrostatic force between the positively charged nucleus and the negatively charged electron that holds the latter in a bound state in the circular orbit. This force provides for the centripetal acceleration of the electron, and one has

$$\frac{mv^2}{r} = \frac{e^2}{4\pi\epsilon_0 r^2}, \quad (16-32)$$

where m stands for the mass of the electron, e for the magnitude of its charge, and ϵ_0 denotes the permittivity of free space.

The *angular momentum* of the electron in the circular orbit is given by

$$L = mvr. \quad (16-33)$$

Finally, the *energy* of the electron, made up of its kinetic and potential energies, is given by the expression

$$E = \frac{1}{2}mv^2 - \frac{e^2}{4\pi\epsilon_0 r}. \quad (16-34)$$

Making use of these relations one can express the energy in terms of the angular momentum as

$$E = -\frac{me^4}{32\pi^2\epsilon_0^2} \frac{1}{L^2}. \quad (16-35)$$

In this classical description the radius of the orbit, and correspondingly the angular

momentum or the energy of the electron, can be given any arbitrarily chosen value. The quantum description of the hydrogen atom, first outlined by Bohr, differs crucially in this respect, as we will see below.

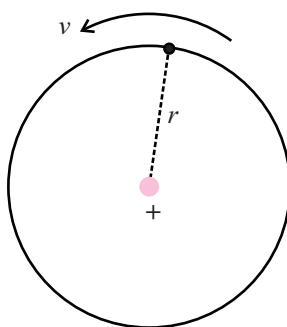


Figure 16-5: A circular orbit of an electron around a positively charged nucleus; the concept of an orbit, however, is of only limited validity in quantum theory.

While the classical description expressed by equations (16-33), (16-34), and (16-35) looks neat, there are lurking anomalies that classical theory failed to resolve. One of these relates to the *stability* of the electron orbit. Electromagnetic theory tells us that an accelerated charged particle radiates energy in the form of electromagnetic waves. Accordingly the electron revolving around the nucleus should continuously lose energy by virtue of its centripetal acceleration, and hence the size of its orbit should go on shrinking till the electron collapses on to the nucleus. In addition to this stability problem, classical theory also failed to explain the *spectrum* of the hydrogen atom, to which we now turn.

16.10.1 The hydrogen spectrum

When gaseous hydrogen is heated to a sufficiently high temperature, it emits electromagnetic radiation that can be analyzed with the help of a spectrometer. One finds that the radiation includes components with a set of characteristic frequencies. Indeed the radiation from any other element like, say, sodium or calcium, is also made up of frequencies characteristic of that element (recall the 'flame test' experiment in your chemistry lab). All the frequencies making up the radiation from a heated material are

said to constitute its *emission spectrum*. In general, the emission spectrum contains, apart from the characteristic frequencies, a set of continuously distributed frequencies as well. This continuous component is absent in the *absorption spectrum* of the material.

When white radiation (made up of all possible frequencies ranging, in principle, from zero to infinity) is passed through cold hydrogen gas, and the resulting radiation is analyzed, it is found that a characteristic set of frequencies is *absent* in the spectrum. These missing frequencies are said to constitute the absorption spectrum of hydrogen, which is simpler than the emission spectrum in that it does not include the set of continuously distributed frequencies.

In a spectrum made up of discrete frequency components, each frequency shows up as a sharp *line* when observed through a spectrometer, and consequently the part of a spectrum made up of a discrete set of frequencies is referred to as a *line spectrum*, as opposed to the *continuous spectrum* (or a *band spectrum*) made up of a continuously distributed set of frequencies.

The characteristic frequencies in the emission spectrum of hydrogen form a discrete set that can be grouped in a number of distinct *series*. Of these, one particular group of frequencies is found to fall in the visible part of the electromagnetic spectrum and is known as the *Balmer series*. While spectroscopists determined the frequencies in the Balmer series and the other parts of the line spectrum of hydrogen, no theoretical explanation could be provided for these frequencies on the basis of the classical theory.

Indeed, classical theory predicts that, as the electron revolves around the nucleus and eventually collapses on to the latter, it should emit radiation made up of only a continuous spectrum, and gives no clue regarding the existence of the line spectrum, of which the Balmer series is a part. It was Bohr who first gave a provisional explanation for the frequencies not only in the Balmer series but in all the other series making up the line spectrum of hydrogen. In this, Bohr introduced the early concepts of quantum theory and initiated the modern approach to the understanding of atomic structure.

16.10.2 Bohr's postulates and the hydrogen spectrum

Bohr based his theory on the circular orbits mentioned above, and retained a number of the classical concepts while at the same time introducing a number of novel postulates in proposing his theory of the structure of the hydrogen atom and of the line spectrum of hydrogen. According to him, the electron can be in any one of a number of *allowed stationary states* corresponding to only a special set of orbits, and no orbit other than these special ones can be occupied by it. He made the extremely important suggestion that the criterion based on which these special orbits are selected out from all the possible orbits that can arise in the classical theory, relates to the *angular momentum* (L) of the electron about the nucleus. More specifically, only those orbits correspond to the stationary states of the electron for which the angular momentum L is an *integral multiple of \hbar* ($\equiv \frac{h}{2\pi}$) :

$$L = n\hbar, \quad (n = 1, 2, 3, \dots). \quad (16-36)$$

Thus, each stationary state is characterized by an integer n , referred to as its *quantum number*. It is similar to the quantum number characterizing the stationary states and energy levels of the harmonic oscillator.

Making use of this postulate introduced by Bohr and of the equation (16-35) relating to the allowed electron orbits one arrives at the following expression for the energy of the stationary state with quantum number n :

$$E_n = -\frac{me^4}{8\epsilon_0^2 h^2} \frac{1}{n^2}. \quad (16-37)$$

In this context, the constant $\frac{me^4}{8\epsilon_0^2 h^3 c}$ is referred to as the *Rydberg constant*. Denoting this by the symbol R one arrives at the compact expression

$$E_n = -hcR \frac{1}{n^2}. \quad (16-38)$$

Thus, the possible energy values of the stationary states of the hydrogen atom are *quantized* along with the angular momentum values. This is so because the quantum number

n can assume only positive integral values. For instance, there cannot be any energy value between $-\frac{hcR}{4}$ and $-\frac{hcR}{9}$ because the quantum number cannot have any value in between 2 and 3. The state with quantum number $n = 1$ has the lowest energy and is known as the *ground state* of the atom. The states with quantum numbers $n = 2, 3, \dots$ are referred to as the first excited state, the second excited state, and so on.

The discrete set of energy values is bounded on the upper side by the energy $E = 0$ (corresponding to $n \rightarrow \infty$). The latter means that the electron just ceases to be bound by the attractive force of the nucleus. There exist states of the electron with energy $E > 0$, but all these states correspond to unbounded motion of the electron, i.e., a motion where the electron moves from infinite distance up to the nucleus and then moves away again to an infinite distance, much like a comet coming up close to the earth and receding away. The possible energy values of all these unbound states are distributed continuously from zero to infinity, but are not relevant so far as the hydrogen atom is concerned because the latter is a bound configuration made up of the nucleus and the electron. The discrete and continuous sets of possible energy values are shown in fig. 16-6.

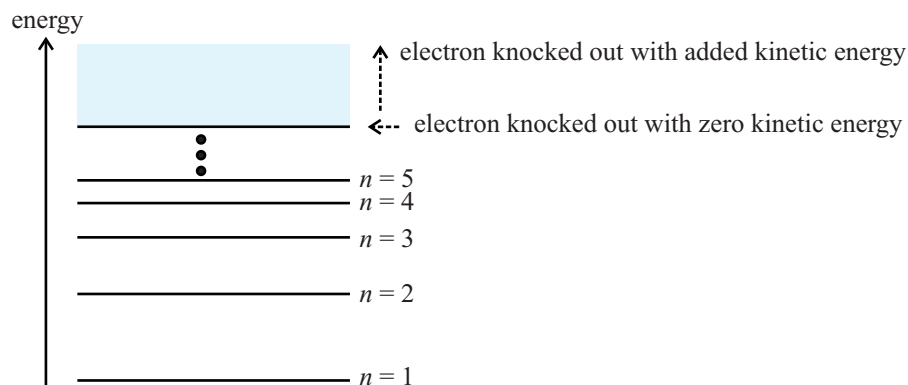


Figure 16-6: Possible energy values of an electron in the Coulomb field of a positive charge (a proton in the present consideration); the vertically upward direction is that of increasing energy; the discrete energy values are indicated with horizontal lines; the continuous set of energy values is indicated with a shading; the two sets are separated by the energy $E = 0$; the lowest energy level corresponds to the ground state; the energy levels with $n = 1, \dots, 5$ are shown schematically; an electron, knocked out of the binding influence of the proton can have any energy from 0 to ∞ , where the zero of the energy scale is assumed to correspond to an electron at an infinite distance from the proton, having zero kinetic energy.

The fact that there exist two sets of states, namely the bound and the unbound ones, is common to the classical and the quantum descriptions, as is the fact that the energy values of the unbound states are distributed continuously. What distinguishes the quantum description is the crucial result that the energies of the possible *bound* states, i.e., the stationary states of the hydrogen atom are quantized, forming a discrete set.

Since Bohr's theory talks of electronic orbits and of stationary states, thereby making use of older classical concepts and the novel quantum ones at the same time, it could not give a complete account of the hydrogen atom including all its structural features. A complete account emerged through a series of subsequent efforts when the quantum theory was developed to the full, superseding completely the classical theory.

The stationary states of Bohr's theory correspond to a special set of circular orbits of the classical theory, each corresponding to a specific value of the quantum number n . The radius of the orbit for quantum number n , and the speed of the electron in this orbit are given by the expressions

$$r_n = \frac{\epsilon_0 h^2}{\pi m e^2} n^2, \quad (16-39a)$$

$$v_n = \frac{e^2}{2\epsilon_0 h} \frac{1}{n}. \quad (16-39b)$$

Problem 16-8

Find the energy required to raise an electron from the ground state with quantum number $n = 1$ to the state with quantum number $n = 3$ in a hydrogen atom, and also the minimum energy required to release the electron from the state $n = 3$ to an unbound state with zero energy.

Answer to Problem 16-8

HINT: The energy of an electron in a hydrogen atom in a state with quantum number n , with reference to a state in which the electron is at an infinitely large distance from the nucleus with

zero kinetic energy, is given by the expression (16-37). Thus, the energy required to raise the electron from the state $n = 1$ to the state $n = 3$ is $E_{1 \rightarrow 3} = \frac{me^4}{8\epsilon_0^2 h^2} \left(\frac{1}{1^2} - \frac{1}{3^2} \right)$, which works out to 1.927×10^{-18} J, i.e., 12.05 eV (approx). The energy required to just release the electron from the $n = 3$ bound state, on the other hand, is $E_{3 \rightarrow \infty} = \frac{me^4}{8\epsilon_0^2 h^2} \left(\frac{1}{3^2} \right) = 1.50$ eV (approx).

16.10.3 Bohr's theory and the quantum theory of the atom

In the complete quantum theory of the hydrogen atom, the classical concept of the electron orbit has no validity. On the other hand, the concept of the stationary states remains central along with the result that the energies of the stationary states are quantized, as are their angular momenta. The expression for the energies of the stationary states coincides precisely with Bohr's expression (16-38), which is why Bohr's theory is hailed as a path-breaking one.

As I mentioned above, the full quantum theoretic description of the hydrogen atom does not allow for the concept of an electron orbit. If this seems a bit strange to you, I suggest you shut off from your mind the picture of the electron as something like a miniature billiards ball (no one 'understands' quantum theory the way one would like to!). Instead, a more appropriate picture is that of a smeared cloud-like distribution, the cloud being relatively more dense at some places and less at some others. The dense regions correspond to those where the electron is more likely to be located compared to the other ones. The classical concept of the electron orbit is, however, not completely devoid of relevance since the expression (16-39a) for the radius of the allowed orbits happens to be related to precisely these regions of high density of the electron cloud.

The full quantum theory of the hydrogen atom is an improvement over Bohr's theory in several respects. One notable aspect of the full quantum theory is that of allowing for a larger class of stationary states, including ones corresponding to the elliptic orbits in the classical theory. Each stationary state is now characterized by a set of *four* quantum numbers rather than one. In addition to the quantum number n mentioned above, referred to as the *principal quantum number* of the state, three more quantum numbers,

termed the angular momentum quantum number (l), the azimuthal quantum number (m_l), and the spin magnetic quantum number (m_s) are needed to completely describe a stationary state. Of these, the spin magnetic quantum number relates to an *internal* characteristic of the electron namely the *spin*.

While I will have more to say about these in chapter 18, one crucial result of the full quantum theory is that the *energy* of the stationary state is determined, in an approximate sense, *solely* by the principal quantum number n , and the expression for the energy is precisely the one (formula (16-38)) given by the Bohr theory.

Accordingly, the principal quantum number n is said to determine the *shell* to which the electron belongs. Each shell can accommodate a certain number of stationary states, characterized by the other quantum numbers mentioned above. All the electrons belonging to any particular shell are characterized by approximately the same energy value while those belonging to some other shell are clearly demarcated in terms of the energy. The shells with $n = 1, 2, 3, \dots$, are referred to respectively as K, L, M, \dots shells.

An important observation to make here relates to the distinction between the wave functions corresponding to the *bound* and *unbound* states of the electron (refer to sections 16.5.1 and 16.5.2).

The state of the free particle mentioned in section 16.2.1 is an example of an unbound state because the free particle is able to move to infinite distances. As we saw in section 16.2.2 this state is represented by a wave function that corresponds to a plane progressive wave. A bound state like that of the electron in the hydrogen atom, on the other hand, means that the electron has no chance of moving away to infinite distances, and such a state is represented by a wave function corresponding to a *standing wave*. Recall that a standing wave is characterized by nodes and antinodes corresponding to regions of maximum and minimum amplitudes. These regions, in turn, correspond to the regions of high and low densities of the electron cloud I mentioned above.

In the quantum theory of atoms, the stationary states, *without reference to the spin*

magnetic quantum number, are sometimes referred to as *orbitals*, recalling thereby the classical picture of electron orbits. Indeed, as we have seen above, electron orbits do retain a measure of relevance in quantum theory and are useful in classifying and describing the stationary states of electrons in an atom.

16.10.4 The hydrogen spectrum: mechanism

Now, then, let us recall the basic results relating to the hydrogen atom. The electron in a hydrogen atom can be in one of many states, of which the stationary states form an important group. A stationary state is characterized by a specific value of energy, which depends on the quantum number n and is given by the expression (16-38). Another important characteristic of the stationary state is that, if the hydrogen atom is left to itself, the state *remains unchanged with time* (by contrast, any *other* state goes on changing or evolving with time).

However, if the atom is acted upon by an external influence then a stationary state of the atom, considered in isolation, can no longer remain unchanged with time. Instead, the atom now acquires the tendency of *making transitions from one stationary state to another*. Consequently, the quantum number as also the energy of the atom can change from time to time.

One such external influence can be an electromagnetic field interacting with the atom. The electromagnetic field can be looked upon as being made of energy quanta or *photons* (see sections 16.9 and 16.13.2 for an introduction to the idea of photons). As the energy of the atom changes during a transition from one stationary state to another, a compensating change then takes place in the energy of the electromagnetic field by way of photons being emitted or absorbed by the atom, whereby the total energy of the atom *and* the electromagnetic field taken together remains conserved.

An interesting observation to make is that the electromagnetic field under consideration *need not be created by charges or currents external to the atom* since, according to quantum mechanical principles, an ubiquitous *vacuum field* always remains even in the absence of photons (refer to sec. 15.6.1). The effect of *this* field is not felt if the atom is

in the ground state ($n = 1$, refer to equation (16-38)). However, if the atom happens to be in an excited state ($n > 1$), it can then emit a photon to change over to *a state with a lower energy*.

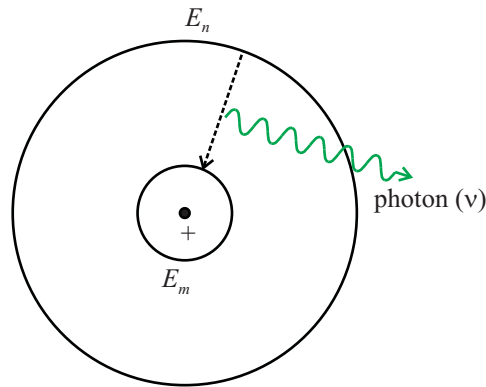


Figure 16-7: Schematic representation of a transition from one stationary state to another where a photon of frequency ν is emitted; E_n and E_m denote the energies of the two stationary states.

Recall from sec. 16.9 that an electromagnetic field can be looked upon as a collection of harmonic oscillators in various different modes - while the modes correspond to stationary waves in an enclosure in the case of black body radiation, traveling wave modes in open space are also possible. An oscillator mode of frequency ν can have quantized energy values given by $E_n = h\nu(n + \frac{1}{2})$ ($n = 0, 1, 2, \dots$) where a quantum number n corresponds to a state of the field that can be described in terms of n photons, each of energy $h\nu$. However, even for photon number zero, the mode possesses an energy $\frac{1}{2}h\nu$, i.e., the ground state energy of the oscillator, which means that there is a non-zero electric and magnetic disturbance in space. Considering the ground states of all the possible modes making up the field, one gets a state of the field in which the photon number is zero, but there is still a non-zero field energy, and hence non-zero electric and magnetic field disturbances everywhere in space. This is commonly referred to as the *vacuum field*, and can have observable effects such as the *spontaneous transition* of an atom from a higher to a lower energy state where, in the initial atom-field state, the photon number is zero for all the field modes, but still the effect

of the vacuum field is to cause the atom to make a transition, emitting a photon.

Suppose then that the atom emits a photon of frequency ν in making a transition from a stationary state with quantum number n and energy E_n to one with quantum number m and energy E_m , where $E_m < E_n$ (see fig. 16-7 for illustration). The principle of conservation of energy requires that the energy emitted in the form of the photon must then be the same as the decrease in energy of the atom:

$$h\nu = E_n - E_m, \quad (16-40a)$$

or, in other words, the frequency of radiation emitted in the transition from a state n to a state m ($m < n$) is given by

$$\nu = \frac{E_n - E_m}{h}. \quad (16-40b)$$

This was one of the derivations of Bohr that made his theory so famous. It explained nicely the observed spectrum of the hydrogen atom. In particular, if the atom is excited to any state with quantum number $n > 2$, and the atom then makes a transition to $n = 2$, then the frequency of the emitted radiation will be

$$\nu = cR\left(\frac{1}{4} - \frac{1}{n^2}\right). \quad (16-41)$$

This expression is found to agree with the observed frequencies of the Balmer series of the hydrogen spectrum.

16.11 Applications of Bohr's theory

What was impressive about Bohr's theory is that it used just a few simple new ideas while still speaking in terms of the familiar classical picture of electron orbits in deriving a number of results that agreed very well with experimental observations and measurements, and solved what had earlier seemed to be deeply puzzling problems in spectroscopy and atomic structure.

Bohr's approach was pursued by Sommerfeld who once again based himself on classical concepts while at the same time making use of new ideas relating to quantization of values of dynamical variables. He included elliptic orbits in his derivation and arrived at a more complete classification of the stationary states, thereby taking Bohr's ideas to their logical conclusion. This made possible the subsequent developments in quantum theory, including a logically sound theory of atomic structure.

Bohr's work, together with experimental investigations by Rutherford and his co-workers, yielded for the first time a reasonably clear and useful picture of the atom (see chapter 18 for further elaboration), commonly referred to as the *planetary model*. This model, together with Bohr and Sommerfeld's theory of stationary states of the hydrogen atom, opened the way to a classification of stationary states of atoms *with more than one electrons* around the nucleus.

This theory of the atomic structure, which preceded the more complete quantum mechanical theory to be developed later, made clever use of results relating to the hydrogen atom. Each of the extra-nuclear electrons was thought to be moving around the nucleus having a certain *effective charge*, the latter reflecting in an approximate sense the effect of the *remaining* electrons in the atom. Having identified the over-all stationary states of the atom in terms of the states of the individual electrons, the *transitions* between the various possible stationary states were then cataloged, thereby accounting for the spectra characterizing the atoms. While this program is not as simple to carry out as these few lines make it appear, a relatively simple instance is provided by the *X-ray spectra of heavy elements*. This I will come back to in chapter 18.

16.12 Bound and unbound systems: standing and traveling waves

We have seen that a stationary state of a free particle, which is an unbound system, is described in terms of a wave function representing a plane monochromatic traveling wave (refer to eq. (16-4)). What, then, does the wave function corresponding to a station-

any state of a *bound* system look like? We have, till now, encountered two such bound systems, namely, the harmonic oscillator, and the electron in the hydrogen atom. The mathematics relating to the wave functions of either of these two systems is a bit involved, but the end-result can be described qualitatively in a simple manner: *the wave function corresponding to a stationary state of a bound system represents a standing wave.*

Standing waves are encountered in numerous situations in classical physics. For instance, examples of standing waves are found in an acoustic wave set up within a pair of reflecting walls, or in the transverse vibrations of a string stretched between fixed ends. These involve stationary patterns made up of nodes and antinodes where there is a periodic variation of the wave function (the excess pressure or the lateral displacement of a point on the string) at each point of a region of space, but the wave as a whole does not propagate.

A characteristic feature of a standing wave is the *discreteness* of the possible values of the frequency characterizing the wave. For instance, the transverse vibrations of a stretched string occur with frequencies which are integral multiples of a certain *fundamental* frequency (refer to section 9.16.1). The discreteness of the possible energy values of a bound system is of an analogous nature.

In other words, the discreteness of the stationary states of a harmonic oscillator or of a hydrogen atom stems from a combination of two basic facts, namely, (i) that the states are described not in the manner of classical physics but in terms of wave functions that satisfy a certain wave equation, and (ii) that the wave functions of a bound system correspond to standing waves.

16.13 Photoelectric effect: Einstein's theory

When electromagnetic radiation of appropriate frequency is made to hit the surface of a metal like, say, sodium, electrons are emitted from the metal. This phenomenon of emission of electrons from certain materials (which include a number of metals and

semiconductors) by electromagnetic radiation is referred to as *photoelectric effect*. This effect can be demonstrated and studied with the help of a set-up like the one shown in figure 16-8.

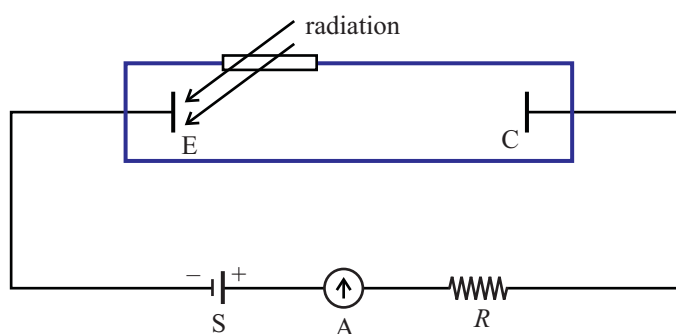


Figure 16-8: Set-up to study and analyze photoelectric effect; E is the emitting surface while C is the collecting electrode; A is a current-measuring device; S is a DC voltage source whose polarity can be reversed; R denotes a resistor; the actual circuit may not be as simple as shown here.

A metallic emitting electrode (E) and a collecting electrode (C) are enclosed in an evacuated chamber in which a window admits electromagnetic radiation of appropriate frequency to fall on E. A circuit made up of a source of EMF (S), a resistor (R), and a sensitive current-meter (A) is established between E and C. The polarity of S can be changed so that C can be either at a higher or a lower potential with respect to E.

16.13.1 Features of photoelectric emission

This arrangement can be used to record a number of interesting features of photoelectric emission. If, for a given intensity of the incident radiation, the potential (V) of C with respect to E is positive then all the electrons emitted from E are collected by C, and A records a current (I). This current remains almost constant when V is increased because all the photoelectrons are collected by C whenever V is positive. This is known as the saturation current for the given intensity of the incident radiation.

However, this entire phenomenon of a current being recorded due to the emission of photoelectrons from E is dependent on the *frequency* (ν) of the radiation. If the frequency is sufficiently low then photoelectric emission *does not occur*, and no photo-current is

recorded. For the time being, we assume that the frequency is high enough for photoelectric emission to take place, and refer back to figure 16-8. If, holding the frequency and intensity of the radiation constant, one now reverses the polarity of S, and records the photocurrent with increasing magnitude of V , one finds that the photo-current persists but decreases gradually till it becomes zero for a value $V = -V_s$ of the potential of C with respect to E. The magnitude (V_s) of V for which the photocurrent becomes zero is termed the *stopping potential* for the given frequency of the incident radiation. This is shown graphically in fig. 16-9.

The lower of the two curves shown in figure 16-9 describes this variation of I with V for a given intensity (J_1) of the incident radiation, the frequency ν being also held constant at a sufficiently high value.

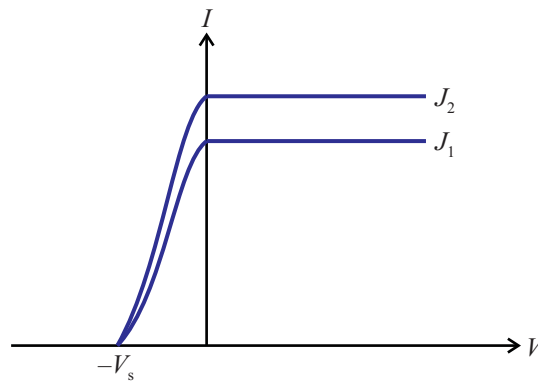


Figure 16-9: Graphical representation of the characteristics of photoelectric emission; variation of photocurrent I with applied voltage V is shown for two values of intensity of radiation, J_1 and J_2 ($J_2 > J_1$) while the frequency ν is held constant; the stopping potential V_s is independent of intensity.

If, now, the experiment is repeated for some other value, say, J_2 , of the intensity of radiation, then one obtains a similar variation, as in the upper curve of fig. 16-9, but with a different value of the saturation current, the latter being higher for $J_2 > J_1$. However, *the stopping potential does not depend on the intensity* since, as seen in figure, both the curves give the same value of the stopping potential.

On the other hand, if the experiment be repeated with various different values of the

frequency, keeping the intensity fixed, one finds that the stopping potential increases with frequency (figure 16-10). One finds that, if the frequency is made to decrease, the stopping potential decreases to zero at some finite value (say, ν_0) of the frequency. This value of the frequency (ν_0) is found to be a characteristic of the emitting material and is referred to as the *threshold frequency* of the latter. Indeed, no photo electric emission from the material under consideration can take place unless the frequency of the incident radiation is higher than the threshold frequency. Moreover, for $\nu > \nu_0$, photoelectric emission does take place for *arbitrarily small values of the intensity*. The effect of lowering the intensity is simply to decrease the photo-current, without stopping the emission altogether.

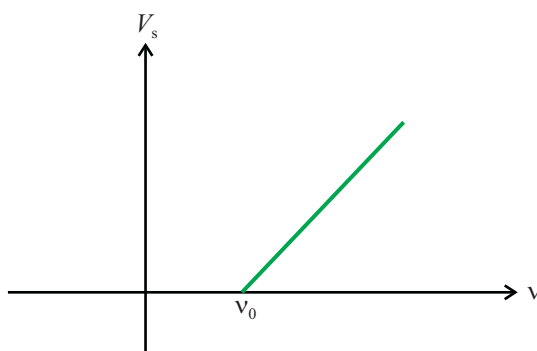


Figure 16-10: Variation of stopping potential with frequency; no photoelectric emission takes place if the frequency is less than the threshold value ν_0 , however large the intensity may be.

16.13.2 The role of photons in photoelectric emission

All these observed features of photoelectric emission could not be accounted for by the classical theory. For instance, classical theory tells us that whatever be the frequency, photoelectric emission should occur if the intensity of radiation be high enough since, for a high intensity of radiation, electrons within the emitting material should receive sufficient energy to come out, overcoming their binding force.

It was Einstein who first gave a reasonably complete account of the observed features of photoelectric effect by invoking the idea of the photon as a quantum of energy, as introduced by Planck in connection with his derivation of the black body spectrum

formula (16-27).

While the photons in the black body radiation were the energy quanta associated with standing wave modes, similar considerations apply to propagating radiation as well. Indeed, the components of electric and magnetic field intensities of propagating monochromatic electromagnetic radiation vary sinusoidally with time, and once again a propagating mode of the field can be looked upon as a quantum mechanical harmonic oscillator of frequency, say, ν . The minimum value by which the energy of the radiation can increase or decrease is once again $h\nu$, and this increase or decrease can once again be described as the appearance or disappearance of an energy quantum, or a *photon*, of frequency ν . Such a photon associated with a progressive wave mode, moreover, carries a momentum just like any other particle such as an electron (by contrast, an energy quantum of black body radiation carries no net momentum). The expressions for energy and momentum of a photon of frequency ν are the de Broglie relations by now familiar to us:

$$E = h\nu, \quad p = \frac{h}{\lambda}, \quad (16-42)$$

where λ stands for the wavelength of the propagating monochromatic radiation and where only the magnitude of the momentum has been considered.

When monochromatic radiation of frequency ν is made to be incident on the surface of a metal or a semiconductor, photons of the same frequency interact with the material and some of these exchange energy with the electrons in it. This can be interpreted as collisions between the photons and the electrons, where the energy of the photon engaged in a collision is transferred to the electron. This energy transfer may be sufficient to knock the electron out of the material, which is how photoelectric emission takes place.

16.13.3 Bound systems and binding energy

A metal or a semiconductor is a *crystalline* material where a large number of atoms are arranged in a regular periodic structure. Electrons in such a material are *bound with the entire crystalline structure*. In this context, it is important to grasp the concept of

a bound system. For instance, a small piece of paper glued on to a board makes up a bound system, and it takes some *energy* to tear the piece of paper away from the board. If the energy of the system made up of the paper separated from the board be taken as zero (in the process of energy accounting, any one energy can be given a pre-assigned value, since energy is undetermined to the extent of an additive constant), and if the energy required to tear the paper apart be E , then the principle of conservation of energy tells us that the energy of the bound system with the paper glued on to the board must have been $-E$ since the tearing energy E added to this initial energy gives the final energy 0.

As another instance of a bound system, consider a hydrogen atom made up of an electron 'glued' to a proton by the attractive Coulomb force between the two. Once again, it takes an energy to knock the electron out of the atom, thereby yielding an unbound electron separated from the proton. The energy of the separated system, with both the proton and the electron at rest, is taken to be zero by convention, in which case the energy of the bound hydrogen atom with the electron in the n th stationary state is given by the expression (16-38). Notice that this energy is a negative quantity, which means that a *positive* energy of equal magnitude is necessary to tear the electron away from the proton. This process of knocking an electron out of an atom is known as *ionization*. It can be accomplished with the help of a photon which supplies the necessary energy to the electron, and the process is termed *photo-ionization*.

In an exactly similar manner, a hydrogen *molecule* is a bound system made up of two protons and two electrons. Looking at any one of these electrons, one can say that it is not bound to any one of the two protons but to the pair of protons *together*. Indeed, the two electrons are shared by the pair of protons and form what is known as a *covalent bond* between the protons. Once again, it takes some energy to knock any one of these electrons out of the hydrogen molecule.

The *minimum energy* necessary to separate the components of a bound system is termed its *binding energy*. On receiving this amount of energy, the components get separated from each other, *without acquiring any kinetic energy* in the separated configuration. If

the bound system receives an amount of energy greater than the binding energy, then the extra amount goes to generate kinetic energy in the components. In this context, an interesting result relates to the situation when one of the components happens to be much lighter than the other. In this case, the extra energy is used up almost *entirely* as the kinetic energy of the *lighter* component.

Incidentally, when I speak of a bound system, I tacitly imply that it is to be looked at as a system made of *two* components. The *same* system may be looked at as one made up of more than two components as well. For instance, in the example of the piece of paper glued on to the board, the components I have in mind are the paper and the board. But, given a sufficient supply of energy, the board can also be broken up into two or more pieces and then one would have to think of a system made up of more than two components. Indeed, the board and the piece of paper are made up of a large number of molecules and the molecules can all be torn away from one another. Similarly, all of the two electrons and the two protons making up the hydrogen molecule can be pulled away from one another, for which a different amount of energy would be required as compared to the energy required to yield just one electron separated from a H_2^+ ion. This latter we term the binding energy of the electron in the hydrogen molecule.

You can now extend your picture of the hydrogen molecule to a molecule made up of more than two atoms where the electrons would be shared between and bound to all the atoms at a time. Or you can even extend your imagination to form an idea as to how the electrons in a crystal lattice are shared between *all the atoms making up the lattice*. Similar to the hydrogen molecule, a certain minimum supply of energy (call it W) would be necessary to knock an electron from the lattice without any kinetic energy being imparted to either of the two separated components (the electron and the rest of the lattice, which is assumed to remain intact). If the energy supplied exceeds W , the extra energy goes to impart kinetic energy to the electron which is by far the lighter of the two separated components. In photoelectric emission, this supply of energy comes from a photon hitting the crystalline material from which the emission takes place.

16.13.4 The basic equation for photoelectric emission

One can now piece together the basic ideas outlined above to arrive at Einstein's explanation of photoelectric effect. For this, one has to start from the possible energy values of an electron in a crystal lattice. Recall that the energy scale we are using assigns zero energy to the configuration where the electron is separated from the lattice, with zero kinetic energy. Recall further that, with the electron *inside* the lattice, the minimum energy to be supplied so as to yield the above configuration is W , this being sometimes referred to as the *work function* of the material under consideration. It then follows that the *maximum* possible energy that an electron can have when inside the lattice, is $-W$. Looking at any one of the large number of electrons that are bound in the lattice, its energy E must then satisfy the inequality

$$E < -W. \quad (16-43)$$

Suppose now, that a photon of frequency ν and energy $h\nu$ hits the material under consideration and imparts its energy to the electron. Assuming that this energy is sufficient to knock the electron out of the lattice, the electron will now have energy $E + h\nu$, and this will appear as the kinetic energy (say, T) of the electron flying out of the material:

$$T = E + h\nu. \quad (16-44)$$

Noting from (16-43) that the maximum possible value of E is $-W$, one arrives at the important result that the maximum possible kinetic energy of a photoelectron, for a photon of frequency ν hitting the emitting material, is given by

$$T_{\max} = h\nu - W. \quad (16-45)$$

This equation is important in explaining some of the observed features of photoelectric emission, summarized in section 16.13.1, because T_{\max} relates to the stopping potential V_s .

T_{\max} **and the stopping potential.**

Imagine a reverse voltage $-V$ ($V > 0$) to be applied to the collecting electrode C in figure 16-8 with reference to the emitting electrode E. This would mean that the potential energy of an electron at C is eV (e being the magnitude of the charge of an electron), assuming that the potential energy is zero at E (recall the arbitrariness in potential energy up to an additive constant). Then, an electron emitted from E with a kinetic energy T in the direction of C would reach C with a kinetic energy $T - eV$, provided $T > eV$ (check this out by making use of the principle of conservation of energy) while, for $T < eV$, the electron would be stopped somewhere between E and C.

In other words, for a given reverse voltage of magnitude V , only those electrons can reach C whose kinetic energy exceeds eV . Electrons with kinetic energy $T = eV$ will then be just stopped at the surface of C. But, as we have already seen, the maximum possible value of the kinetic energy of electrons emitted from E is T_{\max} , given by equation (16-45). This means that, for the given material of the emitting electrode E with a given value of the work function W , the collecting electrode C will cease to receive electrons emitted by E when the magnitude of the reverse voltage exceeds the value (say, V_s) given by

$$eV_s = h\nu - W. \quad (16-46)$$

But, according to what I mentioned in section 16.13.1, the minimum magnitude of the reverse voltage for which the set-up of figure 16-8 fails to record a photocurrent is nothing but the stopping potential. In other words, the stopping potential is related to the frequency of the incident radiation and the work function of the emitting electrode by equation (16-46). It is a neat little equation explaining several of the observed features of photoelectric emission. First of all, (16-45) or, equivalently, (16-46) tells you that for the given material of the emitting electrode, photoemission is possible only if $h\nu > W$, i.e., the frequency of the incident radiation exceeds $\frac{W}{h}$, where W stands for the work function of the material. Using the notation $\nu_0 \equiv \frac{W}{h}$, the condition for photo-emission reads

$$\nu > \nu_0, \quad (16-47)$$

where ν_0 is commonly referred to as the *threshold frequency* of the emitting material. If the frequency of the incident radiation be less than the threshold frequency, photoemission cannot take place, *however large the intensity of the radiation may be*. This result, following from the quantum theory of photoelectric effect, agrees with experimental observations on photoelectric emission and contrasts with what the *classical* theory predicts. The latter implies that it should be possible to make photo-emission occur by using radiation of sufficiently high intensity, regardless of the frequency of the radiation. Moreover, (16-46) does not involve the intensity of the radiation and explains why the stopping potential is independent of the intensity, as we saw in figure 16-9.

Equation (16-46) tells us that the graph of stopping potential against frequency has to be a straight line, as in figure 16-10, which is also found to be consistent with experimental observations. The point of intersection of the straight line with the frequency axis gives one the threshold frequency (and hence the work function $W = h\nu_0$) of the emitting material.

All in all, the theory of photoelectric emission briefly outlined above explains several of the experimentally observed features of the process which the classical theory fails to do. This is essentially how Einstein made use of the novel concept of the photon, introduced by Planck, to account for the observed features of photoelectric process. Subsequent developments in the quantum theory of radiation and of electrons in metals and semiconductors gave a more complete account of photoelectric emission, but Einstein's initial contribution remains a milestone in the history of physics.

Later day developments in the theory of photoelectric effect, and the related theory of photo-emission of electrons from atoms, however, have added a twist here. Careful considerations have shown that the features of photoelectric effect outlined above can be adequately accounted for in a theory of 'mixed' character where the states of the electrons are described quantum mechanically while the electromagnetic field is described in classical terms. In other words, the quantum theory of electromagnetic radiation in terms of photons is, in reality, *not necessary* to account for these features. Of course, if the electromagnetic field is described in accordance with quantum theory

then the conclusions drawn from the mixed theory are corroborated, and in this sense, Einstein's explanation of the photoelectric effect was a pointer in the right direction.

The features outlined above, however, do not exhaust all the experimentally observed facts relating to the photoelectric effect. A number of fine-tuned experiments reveal certain features that can be explained *only* if the quantum description of the electromagnetic field in terms of photons is invoked. In this sense also Einstein's approach of making use of the quantum nature of the radiation field in explaining photoelectric phenomena was a fruitful one.

Problem 16-9

The minimum energy to be supplied to an electron in a metal so as to bring it out is 2.28 eV. Find the threshold frequency, and the corresponding wavelength, for photoelectric emission from the metal. What will be the maximum energy, and the corresponding momentum, of photoelectrons ejected from the metal by light of wavelength 450 nm?.

Answer to Problem 16-9

HINT: The work function for the metal is $W = 2.28 \times 1.6 \times 10^{-19}$ J. Hence, the threshold frequency is $\nu_0 = \frac{W}{h} = \frac{2.28 \times 1.6 \times 10^{-19}}{6.626 \times 10^{-34}} \text{ s}^{-1}$, i.e., $5.51 \times 10^{14} \text{ s}^{-1}$. The corresponding cut-off wavelength is $\lambda_0 = \frac{c}{\nu_0} = 544.5 \text{ nm}$ (approx). When light of wavelength $\lambda = 450 \text{ nm}$ is used, the energy imparted to an electron in the metal by a photon is $E = \frac{hc}{\lambda} = \frac{6.62 \times 10^{-34} \times 3 \times 10^8}{450 \times 10^{-9} \times 1.6 \times 10^{-19}} \text{ eV} = 2.76 \text{ eV}$. Of this, 2.28 eV goes to bring the electron out of the metal, overcoming its work function. Hence the energy of the ejected electron is $E' = E - W = 0.48 \text{ eV} = 0.48 \times 1.6 \times 10^{-19} \text{ J}$, i.e., $7.68 \times 10^{-20} \text{ J}$. An electron of such low energy follows the rules of non-relativistic mechanics (by contrast, an electron of energy of the order of tens of keV, or higher, follows the rules of relativistic mechanics). Hence, its momentum is given by $p = \sqrt{2mE'}$, where $m = 9.1 \times 10^{-31} \text{ kg}$ is the mass of an electron. In other words, the required momentum is $p = 3.74 \times 10^{-25} \text{ kg}\cdot\text{m}\cdot\text{s}^{-1}$.

16.14 The Compton effect

As we have seen, states of the electromagnetic field can be described in terms of photons. While black-body radiation consists of standing wave modes of the electromagnetic field, radiation propagating through space corresponds to traveling wave modes. In the quantum description of the electromagnetic field, each such traveling wave mode appears as a harmonic oscillator, and the stationary states of the oscillator, with a frequency, say, ν can be represented in terms of photons of the same frequency, the energy of a photon being $h\nu$. Moreover, just as a traveling electromagnetic wave causes a transport of energy, it also results in a transport of *momentum* which can be described in terms of photons by saying that a photon, in addition to possessing energy, also carries momentum. The energy and momentum of a photon of frequency ν are given by

$$E = h\nu, \quad p = \frac{h\nu}{c}, \quad (16-48)$$

where p stands for the magnitude of the momentum.

These can be interpreted as the energy and momentum of a particle of rest mass zero.

In chapter 3 I brought up the distinction between the relativistic and non-relativistic points of view in mechanics. The relativistic point of view becomes essential for a particle moving with a velocity close to c , the velocity of light in vacuum, which is often found to be the case for microscopic particles like electrons and protons. In relativistic mechanics, one has to distinguish between the *rest mass* (say, m_0) of a particle and its *moving mass* where the former is the mass observed in a frame of reference in which the particle is at rest while the latter is the mass in a frame in which the particle is moving with, say, a velocity v (refer to chapter 17 for basic ideas in relativistic mechanics). The relation between the energy and momentum of the particle is no longer the same as the one ($E = \frac{p^2}{2m}$) that we have come across in non-relativistic mechanics, but is given by $E = (p^2 c^2 + m_0^2 c^4)^{\frac{1}{2}}$. What is of relevance in this context is to note that, for a particle of rest mass zero, the relation simplifies to $E = pc$, which is exactly what one gets from relations (16-48).

In other words, from the point of view of mechanics, the photon is a particle of zero rest mass where, at the same time, the relations (16-48) point to the fact that it is a quantum of the electromagnetic field.

A photon can transfer part of its energy and momentum to another particle in a process resembling the *elastic collision* between two particles one encounters in mechanics, while *inelastic collision* processes are also possible.

The fact that a photon is a particle with zero rest mass with energy and momentum as in (16-48), is corroborated in *Compton effect*, which is essentially an elastic collision between a high energy photon and an electron. The high energy photon belongs to the frequency range of X-rays or gamma rays, the latter being radiation emitted from atomic nuclei. When a photon with such high energy hits an atom, it transfers part of its energy and momentum to an electron in the atom, thereby knocking it out with some velocity, being itself *scattered* away from the atom with an altered energy and momentum and hence, with an altered frequency. Strictly speaking, a process of this kind is an inelastic collision since part of the energy lost by the photon goes to tear away the electron from the nucleus by supplying its binding energy. But this part of the energy loss of the photon constitutes so small a fraction of its total energy that it can be ignored altogether.

The process of scattering of the photon can then be treated as an elastic collision where the principle of conservation of energy implies that the energy of the incident photon equals the sum of the energy of the scattered photon and the *kinetic energy* of the electron knocked out by the photon (referred to as the *recoil* electron in the scattering process), without regard to its binding energy. At the same time, the collision satisfies the principle of conservation of momentum, where the momentum of the incident photon equals the vector sum of the momentum of the scattered photon and the momentum of the recoil electron.

Fig. 16-11 shows schematically the collision process where a photon of frequency ν comes and hits an electron at rest, its direction of approach being along the line AB.

As a result of the collision, the photon gets scattered at an angle θ to this direction, with an altered energy, which also means an altered frequency, say, ν' , and the recoil electron moves off at an angle ϕ to the direction AB with a momentum, say, p_e , and an energy E_e . In drawing fig. 16-11, I have made use of the fact that the momentum vector of the incident photon, that of the scattered photon, and the momentum vector of the recoil electron have to lie in the same plane (reason out why; consider the plane containing the line of approach of the incident photon and the line along which the photon is scattered; if the line of motion of the recoil electron lies out of this plane, then there is no way momentum can be conserved along a direction perpendicular to it).

The principle of conservation of energy gives a relation between ν , ν' and p_e (E_e and p_e are related by the relativistic energy-momentum relation). Next, the principle of conservation of momentum gives two relations, one for the momentum along the line AB, and the other for the momentum in a perpendicular direction. This gives three equations in all from which one can work out the altered frequency (ν') of the scattered photon, the angle of scattering (θ), and the momentum of the recoil electron (p_e) in terms of the frequency of the incident photon and the angle of recoil (ϕ). A related result gives the change in *wavelength* of the photon in the scattering process as a function of the scattering angle (θ) which I quote below:

$$\Delta\lambda = \frac{h}{m_e c}(1 - \cos\theta), \quad (16-49)$$

where m_e stands for the rest mass of the electron.

This relation, obtained from theoretical considerations, agrees well with observed results in Compton scattering experiments. This, then, can be taken to be another confirmation of the new ideas of quantum theory introduced in the present chapter.

Problem 16-10

Establish the formula (16-49).

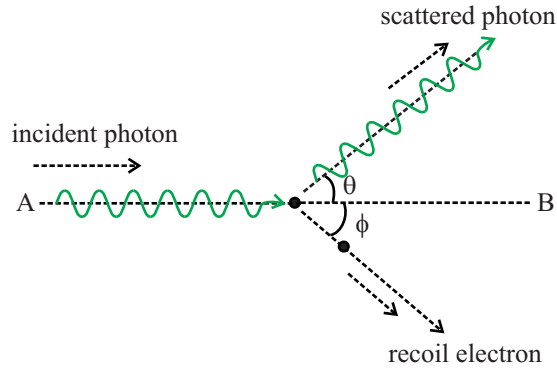


Figure 16-11: Schematic representation of the Compton scattering process; an incident photon with its momentum directed along the line AB hits an electron and gets scattered in a direction making an angle θ while the electron recoils in a direction at an angle ϕ to AB; the process is essentially one of elastic collision between the photon and the electron where the energy lost to overcome the binding of the electron can be neglected; the scattered photon differs in frequency from the incident photon; the change in wavelength is given by eq. (16-49).

Answer to Problem 16-10

HINT: Check that the energy and momentum balance equations mentioned above are as follows:

$$h\nu + m_e c^2 = h\nu' + (p_e^2 c^2 + m_e^2 c^4)^{\frac{1}{2}}, \quad (16-50a)$$

$$\frac{h\nu}{c} = \frac{h\nu'}{c} \cos\theta + p_e \cos\phi, \quad (16-50b)$$

and

$$\frac{h\nu'}{c} \sin\theta = p_e \sin\phi. \quad (16-50c)$$

Making use of these equations, eliminate ϕ and p_e so as to arrive at the formula (16-49), where ν and ν' are to be converted to λ and λ' , the wavelengths characterizing the incident and the scattered photons .

16.15 Quantum theory goes deep: particles and fields

Quantum theory is not just a set of new rules for the description of microscopic particles. It goes deeper than that and opens up a new viewpoint about the particles themselves. It tells us that, in the ultimate analysis, the particles can be identified as quantum states of *fields*.

Fields acquired a new significance within the framework of classical physics itself, with *Maxwell's theory* of the electromagnetic field. That theory gave unequivocal recognition to the fact that the electromagnetic field is a *dynamical system* just like particles or rigid bodies since, analogous to these, the electromagnetic field possesses momentum, energy, and angular momentum of its own, which it can exchange with particles in the course of its interaction with these.

Once the electromagnetic field was recognized as a dynamical system in its own right, the door was opened for the application of quantum rules to describe the behavior of the electromagnetic field, just as quantum rules were applied successfully to the description of particles and their interactions. The first step towards a quantum theory of the electromagnetic field was taken by Planck in his explanation of the black body radiation (see sec. 16.9) where the quantum states of the field were looked upon as the states of so many quantum mechanical harmonic oscillators, each oscillator corresponding to some standing wave mode of the field.

In explaining the black body spectrum, Planck introduced the concept of the photon as the energy quanta of the field, where the number of photons with a given frequency identifies which excited state the corresponding harmonic oscillator is in. However, the photon was not destined to continue in its passive role as just an attribute of the quantum state of the electromagnetic field where it is simply a convenient concept in keeping track of the distribution of the field energy into various possible modes.

As we have seen, the explanation of photoelectric emission by Einstein (sec. 16.13.2) and of the Compton effect (sec. 16.14) gave a firm foundation to the idea that the photon was

itself a particle, capable of exchanging momentum and energy with other bound and free particles.

This is an interesting and novel viewpoint. If photons, which behave as particles in their interaction with other particles such as electrons, are manifestations of the quantum states of a dynamical system at a deeper level, namely, the electromagnetic field, then maybe *other particles* like the electrons and the protons are also similar manifestations of fields, where the latter are the underlying dynamical systems having ultimate significance. One great by-product of this point of view is the recognition that what we perceive as a particle like, say, an electron, is not an immutable entity, but can be *created and annihilated* as the state of the underlying field changes.

Consider once again the photons making up the black body radiation. As the temperature of the black body radiation is made to increase, the modes of the electromagnetic field are excited to higher energy states, corresponding to a larger number of photons making up the radiation. In other words, new photons are created as the state of the field changes where, alternatively, the annihilation of photons is also possible, as when the temperature of the black body radiation is reduced. Photons are also created and annihilated in *atomic transitions* such as in atomic absorption and emission processes (recall, for instance, how these absorption and emission processes are made use of in the production of a laser beam (sec. 15.6)).

If the viewpoint of ascribing a basic significance to the fields as dynamical systems is a valid one, and if the particles, in the ultimate analysis, are just so many quantum states of the fields, then it must follow that the particles are not immutable entities, and that processes of creation and annihilation of the particles should be there as in the case of photons.

Such processes are indeed observed, though not as easily as the ones involving the creation and annihilation of photons. For instance, processes involving the creation and annihilation of electrons and other particles are observed in *cosmic ray* phenomenon, where highly energetic particles entering into the upper layers of the atmosphere from

the outer space interact with other particles, resulting in such processes of creation and annihilation. These processes can thus be described as interactions of the underlying fields as dynamical systems whereby the states of the fields get changed, and this change shows up as a creation and destruction of one or more particles.

The consistent application of the quantum principles to fields led to the development of the theory describing the interaction of the electromagnetic field with charged particles like the electron and the positron, the latter being the *antiparticle* of the electron (see sec. 18.8.9.3). At the same time, a scheme for the classification of the elementary particles and their interactions was developed that lends order to an almost bewildering variety of particles and their interaction processes that were observed in high energy experiments. The theory of these processes is, from a basic point of view, a theory of the underlying set of fields (refer to section 18.8.9 for a brief introduction to the relevant basic concepts).

Thus, from its beginnings as a theory describing the behavior of particles in the microscopic domain, quantum theory has delved into deeper levels of reality and has developed into the quantum theory of fields in which elementary particles and their interactions find a consistent interpretation.

Chapter 17

Relativity: the special and the general theory

17.1 Relativity: Introduction

17.1.1 Introduction: frames of reference, inertial frames

We saw in chapter 3 that every physical observation and every measurement takes place in some *frame of reference* or other. A frame of reference is defined by a set of points and lines rigidly fixed with respect to one another (such as the origin and the axes of a Cartesian co-ordinate system) and, in addition, by a set of clocks fixed in the system. In any given frame of reference, one can make use of a system of co-ordinates so as to specify the spatial location of a point, and a clock to specify the time of occurrence of an *event* at that point. Choosing a different co-ordinate system and, possibly, a different set of clocks in a frame of reference, or going over from one frame to another results in a space-time *co-ordinate transformation*.

A co-ordinate system stands for a consistent way of identifying points and world lines in the four dimensional space-time continuum, where the term ‘world lines’ will be explained in what follows. Briefly, a world line depicts the history of motion of a particle in space and time. A point corresponds to the crossing of two world lines,

and represents an ‘event’. The set of all possible events makes up a four dimensional space, which we refer to as the space-time continuum. It is the set of events and world lines that are to be considered as being of basic physical relevance. The spatial and temporal co-ordinates constitute a set of book-keeping entries for the events in space-time, where there may be more than one (actually, an infinite number of) such sets of entries, which brings in the idea of co-ordinate transformations. However, all this is to anticipate what I am going to tell you in the following sections in this chapter. At times, it helps to have a glimpse of things to follow.

In the *non-relativistic* theory of chapter 3, one commonly makes use of an *inertial* frame of reference since the so-called inertial forces do not appear in the Newtonian equations of motion written in such a frame, as a result of which the equations of motion assume the same form (i.e., are *form-invariant*) in all inertial frames. In other words, in the context of describing and explaining physical phenomena on the basis of the Newtonian equations, all inertial frames are *equivalent*.

17.1.2 Introduction: the Galilean principle of equivalence

This can be seen by looking at the way the Newtonian equations get transformed under a co-ordinate transformation corresponding to the transformation from one inertial frame to another - the so-called *Galilean* transformation (refer to section 3.9.3).

The term ‘co-ordinate transformation’ will mean, in general, a transformation of spatial *and* temporal co-ordinates of events.

Let us use a set of Cartesian variables x, y, z to specify the location of a point in an inertial frame S, where these constitute the components of the position vector \mathbf{r} of the point, and also let the variable t denote the time of occurrence of an event at \mathbf{r} in S. Let the corresponding variables in a second inertial frame S' be, respectively, x', y', z' , and t' , where the former three make up the position vector \mathbf{r}' . If the velocity of S' relative to

S be \mathbf{V} , then the transformation equations read

$$t' = t, \quad \mathbf{r}' = \mathbf{r} - \mathbf{V}t, \quad (17-1)$$

where these transformation equations imply the transformations (3-41), and (3-44), (3-45) (check this out).

In writing the above transformation equations we have assumed that the clocks in S and S' are synchronized, i.e., these are made to agree at any chosen point of time. The first relation in (17-1) will then hold for all times. More generally, t' will differ from t by an additive constant which, however, will have no effect on the transformation of the Newtonian equations of motion.

The statement that all inertial frames are equivalent in describing physical phenomena based on the Newtonian equations of mechanics, is referred to as the *Galilean* (or *non-relativistic*) principle of equivalence.

17.1.3 Introduction: the non-relativistic and the relativistic

What is of importance here is that, in the non-relativistic theory, the velocity \mathbf{V} occurring in the above expression is to be small - small compared to the *velocity of light* in free space because, for relative velocities comparable to the velocity of light, the transformation formulae look different and the Galilean principle of equivalence no longer holds, which is precisely what we will see below as we get into *relativistic* considerations. Further, in describing the motion of a particle in a frame of reference, we can use the Newtonian equations only when the velocity of the particle is small compared to the velocity of light, which constitutes the condition of validity of the non-relativistic approach to the mechanics of the particle. For particle velocities comparable to the velocity of light, the equations of motion are to be modified into the equations of *relativistic mechanics*.

Of equal importance is the fact that, in the relativistic theory, one has to distinguish between situations where gravitational fields are absent (or are negligible) and those

where gravitational fields make their presence felt. These require different approaches in describing physical phenomena - those relating to the *special* and the *general* theories of relativity respectively.

One thus has to reckon with *three* different approaches in describing physical phenomena, consisting of motions of particles or of systems of particles - the non-relativistic approach, the approach of the special theory of relativity, and that of the general theory of relativity. You may possibly be having the impression that these are mutually exclusive ways of describing physical phenomena, depending on the relevant velocities and strengths of gravitational fields involved. If this were the case, physics would have been a confusing affair indeed.

In reality, physics aims at giving us a unitary approach in describing various phenomena, regardless of the values of parameters involved, where seemingly different approaches turn out to be related to one another in definite ways. Thus, the non-relativistic description in an inertial frame can be seen to result from the special relativistic description in the limit of small velocities, while the special relativistic description, in turn, constitutes a limiting instance of the general relativistic one in the limit of infinitesimally weak gravitational fields.

Incidentally, the term 'physical phenomena' is to be interpreted to mean, not just the motions of particles and systems of particles in accordance with Newton's laws but, more generally, *electromagnetic* phenomena as well. In the non-relativistic theory, electromagnetic forces on particles are introduced in specified forms in the Newtonian equations of motion which, strictly speaking, lack consistency since these imply action-at-a-distance. Phenomena involving electromagnetic field variations cannot be described consistently in the non-relativistic theory since the Maxwell equations admit of a consistent interpretation only in the special relativistic description. In other words, the special theory of relativity offers a consistent description of a broader class of phenomena, namely, those involving mechanical motion of particles based on electrical and magnetic forces, as also the ones involving electromagnetic field variations.

In reality, the two types of phenomena mentioned above are not mutually exclusive since the motion of particles under electrical and magnetic forces are mediated by electromagnetic field variations.

Gravitation is accommodated in the non-relativistic theory as a force acting on particles, which again involves action-at-a-distance, implying the necessity of a relativistic approach. However, the consistent way of accommodating gravitation in the relativistic framework differs fundamentally from that adopted in the case of electromagnetic phenomena. More precisely, while electromagnetic phenomena fits in the framework of the special theory of relativity, gravitation requires a broader formalism, namely the one of the general theory of relativity where it appears not as a ‘force’ on particles but as a distortion of the metrical properties of space and time through which the particles move.

17.1.4 Introduction: the equivalence principles

The non-relativistic theory is based on the Galilean principle of equivalence (sections 3.10.2, 17.1.2), which tells us that the Newtonian equations are invariant under the Galilean transformations or, in other words, that all inertial frames are equivalent for sufficiently small relative velocities. The special theory of relativity is developed from a more general principle of equivalence, which states that *all* inertial frames are equivalent, regardless of the magnitudes of the relevant relative velocities.

Here an inertial frame is defined as one in which a free particle, i.e., one not acted upon by any force, moves with uniform velocity. In the pre-general relativistic framework, this means a particle not under the influence of electromagnetic and gravitational forces. However, in the general relativistic framework, gravitation is not included in the category of a force, and so a free particle is one that moves freely in a gravitational field, not acted upon by electromagnetic forces.

For the sake of completeness, one has to broaden the concept of force by way of including the weak and strong forces characterizing the interactions of elementary particles. However, the accommodation of the weak and strong forces necessitates another

round of generalization of the theoretical framework, namely the one of quantum field theory. A complete and consistent field theory accommodating gravitation is, for now, not known. We will leave these out of consideration in this introductory exposition of the general theory of relativity.

With this broadening of the concept of a free particle, one needs a broader view of the principle of equivalence as well, which is precisely the point of departure of the general theory of relativity. We will have a glimpse into this in later sections in this chapter.

17.2 The special theory of relativity

17.2.1 Inertial frames and the velocity of light

The special theory of relativity is, primarily, a restructuring of our idea of space and time that becomes of essential relevance when one considers large velocities of particles and of frames of reference. As I have mentioned above, this means velocities comparable to the speed of light in free space. This speed of light in free space (which is the same for electromagnetic waves of all frequencies) is a fundamental constant of nature and plays a crucial role in the formulation of the theory of relativity. Indeed, if one adopts the view that all inertial frames of reference are to be equivalent (and not just ones with small relative velocities), then a conflict arises in accommodating electromagnetic phenomena in the theory since it turns out that Maxwell's equations are not form-invariant under the Galilean transformations. Instead, if one starts from the assumption that Maxwell's equations are also to be form-invariant then one has to seek a modification in the co-ordinate transformations corresponding to the change from one inertial frame to another. Since Maxwell's equations in free space imply a constant value for the speed of light ($c = \frac{1}{\sqrt{\epsilon_0 \mu_0}}$) it seems natural to assume that the *speed of light is the same for all inertial frames of reference*.

In other words, the point of departure for the special theory of relativity is a broadened formulation of the principle of equivalence (as compared with the Galilean equivalence principle) where one requires that the speed of light in free space is to be the same in all

inertial frames of reference.

17.2.2 The Lorentz transformation formulae

Starting from this broadened principle of equivalence, Einstein arrived at a fundamentally new form for the co-ordinate transformation corresponding to the change from one inertial frame to another. I will now tell you what this transformation looks like.

In this, I will, for the sake of simplicity, consider two inertial frames of reference S and S' as shown in fig. 17-1 where the spatial co-ordinates will be referred to in terms of the Cartesian systems $OXYZ$ and $O'X'Y'Z'$ (see caption to figure), and time co-ordinates of events will be chosen such that the origins O and O' of the two systems coincide at $t = t' = 0$. This corresponds to a convenient choice of the origin of time in each of the two frames. As explained in the caption, the frame S' is assumed to move with a velocity V with respect to S , measured along the positive x -axis of either co-ordinate system.

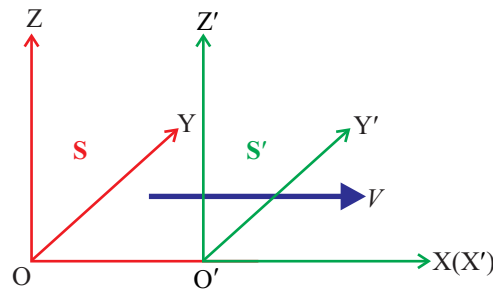


Figure 17-1: Transformation from one inertial frame to another; S and S' are two inertial frames in which Cartesian co-ordinate systems $OXYZ$ and $O'X'Y'Z'$ are chosen with their respective co-ordinate axes parallel, as shown, the axes OX and $O'X'$ being along the same lines; the origins O and O' are assumed to coincide at $t = t' = 0$; the system S' is assumed to move with a uniform velocity V with respect to S , measured along the positive direction of the x -axis; with such a choice of the two inertial frames and of co-ordinates in these, the equations of co-ordinate transformation assume a relatively simple form as in equations (17-2a), (17-2b), differing from the Galilean transformation equations (17-3) for the same two frames.

These assumptions about the two frames of reference are convenient ones but are not too restrictive. With these assumptions, the transformation equations from co-ordinates t, x, y, z to t', x', y', z' describing any arbitrarily chosen event in the two frames appear as:

$$t' = \gamma(t - \frac{Vx}{c^2}), \quad (17-2a)$$

$$x' = \gamma(x - Vt), \quad y' = y, \quad z' = z, \quad (17-2b)$$

where we use the notation

$$\gamma = \frac{1}{\sqrt{1 - \frac{V^2}{c^2}}}. \quad (17-2c)$$

This is to be compared with the Galilean transformation between the same two co-ordinate systems, that reads

$$t' = t, \quad x' = x - Vt, \quad y' = y, \quad z' = z, \quad (17-3)$$

(check that these formulae follow from the general Galilean transformation formula (17-1) for the two frames described in fig. 17-1).

The relativistic transformation formulae between any two inertial frames is referred to as a *Lorentz transformation*. Equations (17-2a), (17-2b) thus describe a special instance of the Lorentz transformation, namely, the one between two frames related as in fig. 17-1. A more general instance can also be constructed, where the Cartesian systems OXYZ and O'X'Y'Z' are arbitrarily oriented with respect to one another and where the velocity of S' relative to S is along any arbitrarily specified direction. I will give you the transformation formulae for such a general situation in sec. 17.2.4.

In the above transformation formulae, t, t' denote the times of occurrence of an event in the two frames under consideration, while x, y, z and x', y', z' denote the spatial co-ordinates of the point where the event occurs. Thus, an event is described in any given frame by *four* co-ordinates, and the transformation relates the four co-ordinates describing an event in one frame to the corresponding four in another.

One fundamental difference between the Galilean and the Lorentz transformation formulae (equations (17-3) and (17-2a)- (17-2c)) is that, in the former, the time co-ordinate remains unchanged (subject to an appropriate choice of the origin of time in each frame; this we will always assume to be the case) while in the latter the time gets transformed in an essential way, where the transformation depends on the spatial co-ordinate(s) of the event. In other words, in the non-relativistic theory, the time co-ordinate has a special status, corresponding to the concept of *absolute time*, independent of the frame of reference of the observer. In the relativistic theory, on the other hand, the time of an event is essentially dependent on the observer's reference frame and, in general, all the four co-ordinates get transformed in a Lorentz transformation in an interdependent manner.

However, this basic difference notwithstanding, the two transformations are related to each other in a certain definite manner: as we will see in the following sections, *consequences of the Lorentz transformation formulae relating to measurements of spatial and temporal intervals reduce to those following from the Galilean transformation in the limit of small velocities*, in which terms of degree two and higher in the ratio $\frac{V}{c}$ can be ignored. It is in this limiting sense that the non-relativistic theory is related to the special theory of relativity that constitutes a generalization of the former (refer to sec. 17.1.3).

Before we continue, I want you to note that the four co-ordinates (t, x, y, z) are not dimensionally homogeneous, since the dimension of t differs from that of x, y or z . A more appropriate choice is one where one uses ct instead of the time co-ordinate t . In this case the Lorentz transformation equations (17-2a)- (17-2c) assume the simpler form

$$\begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix} = L(V) \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}, \quad (17-4a)$$

where, for the sake of convenience, we have used a *matrix* representation, with the

column $\begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}$ representing the four space-time co-ordinates of an event in the frame S,

and the column $\begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix}$ representing the co-ordinates of the same event in S'. In this matrix representation, $L(V)$ stands for the 4×4 *transformation matrix* given by

$$L(V) = \begin{pmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad (17-4b)$$

where the parameter γ is defined as in (17-2c), and β is given by

$$\beta = \frac{V}{c}, \quad (17-4c)$$

(check this out). The two are related as

$$\gamma^2 = \frac{1}{1 - \beta^2}, \quad \beta^2 = 1 - \frac{1}{\gamma^2}. \quad (17-5)$$

Note that in the special case considered here, where the frames S and S' are related as in fig. 17-1, there is no change in the spatial co-ordinates y, z , perpendicular to the direction of the relative velocity, under the Lorentz transform. Hence, in such a special case, it suffices to take into account only two space-time co-ordinates instead of four, these being the co-ordinates ct and x in the case under consideration.

17.2.3 Space-time interval

Consider two events with space-time co-ordinates ct_1, x_1, y_1, z_1 and ct_2, x_2, y_2, z_2 in an inertial frame S. The quantity

$$s^2 \equiv (ct_2 - ct_1)^2 - (x_2 - x_1)^2 - (y_2 - y_1)^2 - (z_2 - z_1)^2, \quad (17-6a)$$

is referred to as the squared *space-time separation* (or, equivalently, the squared space-time *interval*) between the two events. Depending on the two events under consideration, the squared space time separation can be either positive or negative, and its square root (the space-time separation s) can be a real or an imaginary quantity, which is why we will more often refer to the squared space-time separation s^2 rather than the separation itself.

Let the space-time co-ordinates of the same two events, referred to a second inertial frame S', be ct'_1, x'_1, y'_1, z'_1 and ct'_2, x'_2, y'_2, z'_2 respectively. Then the squared space-time separation, from the point of view of the frame S' will be

$$s'^2 \equiv (ct'_2 - ct'_1)^2 - (x'_2 - x'_1)^2 - (y'_2 - y'_1)^2 - (z'_2 - z'_1)^2. \quad (17-6b)$$

As can be checked from the formulae (17-2a)- (17-2c), the squared separations in the two frames *are the same*.

Problem 17-1

Check this statement out.

Answer to Problem 17-1

HINT: The relevant transformation equations for the two events are

$$ct'_1 = \gamma(ct_1 - \beta x_1), \quad ct'_2 = \gamma(ct_2 - \beta x_2), \quad (17-7a)$$

$$x'_1 = \gamma(x_1 - \beta t_1), \quad y'_1 = y_1, \quad z'_1 = z_1; \quad x'_2 = \gamma(x_2 - \beta t_2), \quad y'_2 = y_2, \quad z'_2 = z_2, \quad (17-7b)$$

where γ, β are given by (17-2c), (17-4c), and are related as in (17-5).

The required result is obtained on making use of these formulae in (17-6b).

In other words, the squared space-time separation between the two events, expressed in terms of the space-time co-ordinates of these events relative to two inertial frames, is an *invariant* quantity in a Lorentz transformation:

$$s^2 = s'^2, \quad (17-8)$$

even though the co-ordinates themselves get changed in the transformation.

While you have verified this statement in the special case of two inertial frames S and S' related to each other as in fig. 17-1, it remains true for the transformation between any two arbitrarily chosen inertial frames, when the transformation formulae assume a more general form, as we will see in sec. 17.2.4 below.

Incidentally, given an event with space-time co-ordinates (ct, x, y, z) referred to a frame S, the quantity

$$s^2 \equiv c^2 t^2 - x^2 - y^2 - z^2, \quad (17-9)$$

stands for the squared space-time separation between this event and the event with co-ordinates $(0, 0, 0, 0)$. According to the above result on the invariance of the space-time separation in a Lorentz transformation, we have

$$c^2 t^2 - x^2 - y^2 - z^2 = c^2 t'^2 - x'^2 - y'^2 - z'^2, \quad (17-10)$$

where (ct', x', y', z') stands for the co-ordinate of the event under consideration, referred to the frame S' (recall that, according to our definition of the co-ordinates in S', the event $(0, 0, 0, 0)$ in S has the same co-ordinates in S' as well).

The use of the same symbol s^2 in the two expressions in (17-6a), (17-9) need not cause

confusion, since the meaning in either case is clear from the context. The symbol s will be used generically to denote space-time separations.

Problem 17-2

Consider inertial frames S and S' related to each other as in fig. 17-1, and a third frame S'' attached to a right handed Cartesian co-ordinate system whose axes are all parallel to the corresponding axes of S , S' , and are coincident with the latter at $t = t' = t'' = 0$; the origin of S'' moves with uniform velocity V' with respect to S' along the z -axis of the latter. Find the transformation from space-time co-ordinates (ct, x, y, z) in S of an event to co-ordinates (ct'', x'', y'', z'') of the same event in S'' . Establish the invariance relation $c^2 t''^2 - x''^2 - y''^2 - z''^2 = c^2 t^2 - x^2 - y^2 - z^2$.

Answer to Problem 17-2

HINT: The transformation will be of the form (17-4a), where now only the co-ordinates ct', z' in S' will be transformed to ct'', z'' in S'' , with V replaced with V' (and, correspondingly, β, γ with $\beta' = \frac{V'}{c}, \gamma' = \frac{1}{\sqrt{1 - \frac{V'^2}{c^2}}}$), giving

$$\begin{pmatrix} ct'' \\ x'' \\ y'' \\ z'' \end{pmatrix} = L(V') \begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix}, \quad (17-11a)$$

where

$$L(V') = \begin{pmatrix} \gamma' & 0 & 0 & -\beta'\gamma' \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\beta'\gamma' & 0 & 0 & \gamma' \end{pmatrix}. \quad (17-11b)$$

In this transformation, β', γ' are given by

$$\beta' = \frac{V'}{c}, \gamma' = \frac{1}{\sqrt{1 - \frac{V'^2}{c^2}}}. \quad (17-11c)$$

The resulting transformation from S to S'' is then given by

$$\begin{pmatrix} ct'' \\ x'' \\ y'' \\ z'' \end{pmatrix} = L \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}, \quad (17-12a)$$

where the transformation matrix is now

$$L = L(V')L(V) = \begin{pmatrix} \gamma\gamma' & -\beta\gamma\gamma' & 0 & -\beta'\gamma' \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\beta'\gamma\gamma' & \beta\beta'\gamma\gamma' & 0 & \gamma' \end{pmatrix}. \quad (17-12b)$$

Considering the events (ct, x, y, z) and $(0, 0, 0, 0)$ in the frame S and the same two events in S'', we have, from above,

$$s''^2 = (\gamma\gamma'ct_0 - \beta\gamma\gamma'x - \beta'\gamma'z)^2 - (\beta\gamma ct + \gamma x)^2 - y^2 - (-\beta'\gamma\gamma'ct + \beta\beta'\gamma\gamma'x + \gamma'z)^2. \quad (17-13)$$

On making use of the relations (17-5), and the corresponding relations involving β', γ' , this is seen to give $s''^2 = s^2$, as expected, since S and S'' are both inertial frames.

17.2.4 Lorentz transformation: the general form

The transformation formulae (17-2a)-(17-2c) (or, equivalently, the matrix formula (17-4a), with $L(V)$ and β defined as in (17-4b), (17-4c)) are valid for the special case of two frames defined as in fig. 17-1, where the respective axes of the Cartesian co-ordinate systems chosen in the two frames are parallel to each other, and the relative velocity is along one of the axes. More general forms of the Lorentz transformation arise when the axes and the direction of the relative velocity do not meet these restrictive conditions. The transformation, however can still be represented in the form (17-4a), where the 4×4 transformation matrix $L(V)$ differs from that given in (17-4b).

Fig. 17-2 depicts two inertial frames S, S', with Cartesian co-ordinate systems OXYZ and O'X'Y'Z' chosen in the two frames, where the relative orientation of the two Cartesian

systems, and the direction of the velocity of S' relative to S are fixed arbitrarily. The origin O' of the system in S' moves with velocity V along the dotted line in the direction of the arrow, with respect to O . We assume that the two origins (O, O') coincide with each other at $t = t' = 0$. The relative orientation of the two co-ordinate systems remains unchanged during their relative motion, as S' possesses only a translational velocity with respect to S (since, otherwise, S' would be a non-inertial frame).

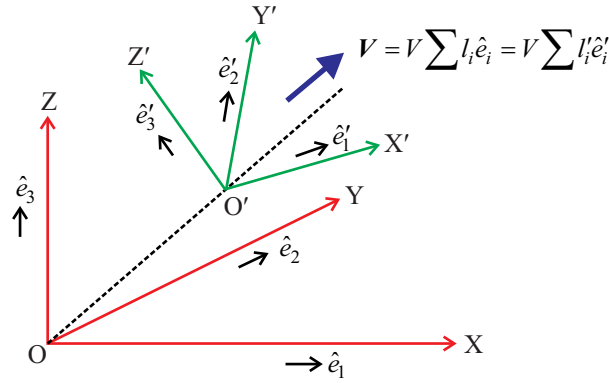


Figure 17-2: Depicting the geometry for a general Lorentz transformation; the Cartesian co-ordinate systems $OXYZ$ and $O'X'Y'Z'$ in frames S and S' are oriented with respect to each other in an arbitrarily chosen manner, with reference to which the orientation shown in fig. 17-1 is a special case; the origins of the two systems are assumed to coincide with each other at $t = t' = 0$; the system S' is assumed to move with a uniform velocity V with respect to S along the dotted line in the direction of the arrow, the said direction being one chosen arbitrarily with reference to either set of co-ordinate axes; once again, the case shown in fig. 17-1 is a special instance; the transformation equations from the space-time co-ordinates ct, x, y, z of an event referred to the frame S to the co-ordinates ct', x', y', z' of the same event referred to S' then constitute the general form of Lorentz transformation (refer to eq. (17-17)), of which the equations (17-2a)-(17-2b) constitute a special instance.

Let the unit vectors along the three co-ordinate axes attached to the frame S be denoted by \hat{e}_i ($i = 1, 2, 3$) while the corresponding unit vectors for S' be \hat{e}'_i ($i = 1, 2, 3$). Further, let the unit vector along the relative velocity of S' with respect to S have direction cosines l_i and l'_i ($i = 1, 2, 3$) with respect to the Cartesian systems in S and S' respectively. The relative orientations of the two co-ordinate systems is completely described by the quantities

$$t_{ij} \equiv \hat{e}'_i \cdot \hat{e}_j, \quad (i, j = 1, 2, 3) \quad (17-14)$$

that constitute the elements of a 3×3 *orthogonal* matrix ($\sum_k t_{ik}t_{jk} = \delta_{ij}$ ($i, j = 1, 2, 3$), where δ_{ij} stands for the Kronecker delta symbol with indices i, j), termed the *transformation matrix* from the co-ordinate systems OXYZ to O'X'Y'Z'. The transformation matrix relates the components, in the two co-ordinate systems, of the unit vector along the direction of the relative velocity shown in fig. 17-2, as

$$l'_i = \sum_j t_{ij}l_j \quad (i = 1, 2, 3), \quad (17-15)$$

(check this out; refer to sec. 2.9).

The change from one Cartesian co-ordinate system to another one with a fixed orientation relative to the former is a transformation in three dimensional space (while a Lorentz transformation describes a change from one frame of reference to another, and describes a transformation in four-dimensional space-time); if (x_1, x_2, x_3) and (x'_1, x'_2, x'_3) denote the co-ordinates of a point in the two co-ordinate systems in the three dimensional space, then

$$x_1^2 + x_2^2 + x_3^2 = x'^2_1 + x'^2_2 + x'^2_3, \quad (17-16)$$

which tells us that the squared spatial separation between the point and the origin (assumed to be the same for the two co-ordinate systems) remains invariant under the transformation. This invariance is analogous to the relation (17-10) that holds for a Lorentz transformation. Referring to fig. 17-2, the two co-ordinate systems do not have the same origin, in which case we have to consider the separation between two given points in space (the co-ordinates $(0, 0, 0)$ in the two frames do not correspond to the same point), analogous to the separation between two space-time events.

With this specification of the two reference frames S and S', the Lorentz transformation from the former to the latter can once again be expressed in the form (17-4a), where

now the transformation matrix $L(V)$ turns out to be

$$L(V) = \begin{pmatrix} \gamma & -\beta\gamma l_1 & -\beta\gamma l_2 & -\beta\gamma l_3 \\ -\beta\gamma l'_1 & (\gamma - 1)l'_1 l_1 + t_{11} & (\gamma - 1)l'_1 l_2 + t_{12} & (\gamma - 1)l'_1 l_3 + t_{13} \\ -\beta\gamma l'_2 & (\gamma - 1)l'_2 l_1 + t_{21} & (\gamma - 1)l'_2 l_2 + t_{22} & (\gamma - 1)l'_2 l_3 + t_{23} \\ -\beta\gamma l'_3 & (\gamma - 1)l'_3 l_1 + t_{31} & (\gamma - 1)l'_3 l_2 + t_{32} & (\gamma - 1)l'_3 l_3 + t_{33} \end{pmatrix}, \quad (17-17)$$

where γ and β are given by (17-2c), (17-4c), as before.

Problem 17-3

Check that, in the special case where the frames S , S' are related as in fig. 17-1, the above transformation matrix reduces to the one in (17-4b).

Answer to Problem 17-3

HINT: In this special case, one has $t_{ij} = \delta_{ij}$, $l_1 = l'_1 = 1$, while all the other relevant direction cosines are zero.

The central result that follows from the general Lorentz transformation matrix (17-17) is that *the squared space-time separation between any two given events remains invariant under the transformation from S to S' as in the special case of the transformation described in fig. 17-1 (check this statement out; a more convenient derivation of this result follows from (17-18) below).*

A compact description of the general Lorentz transformation formula can be given as follows. Let the spatial co-ordinates of the event (ct, x, y, z) in S make up the vector \mathbf{r} , while the spatial co-ordinates of the same event in S' make up the vector \mathbf{r}' . For either of these two vectors, we denote the component along the direction of the relative velocity by the suffix ' \parallel ', and the two dimensional vector perpendicular to the direction of the relative velocity by the suffix ' \perp '. Then, an examination of the transformation (17-2a), (17-2b) suggests the following form of the transformation for the

general case:

$$r'_{\parallel} = \gamma(r_{\parallel} - \beta ct), \quad \mathbf{r}'_{\perp} = \mathbf{r}_{\perp}, \quad ct' = \gamma(ct - \beta r_{\parallel}). \quad (17-18)$$

Expressing this in terms of the space-time co-ordinates in the frames S and S' related as in 17-2, one arrives at the transformation formula (17-4a) with the Lorentz transformation matrix $L(V)$ given by (17-17). However, I skip the derivation here since it requires a bit of involved algebra, and does not lead to new concepts.

Note that, in this general case, all the four space-time co-ordinates of an event get transformed simultaneously, in contrast to the special case for which the transformation matrix is given by (17-4b), where two of the spatial co-ordinates remain unchanged in the transformation. One need not, however, put too much of significance in it since, in the compact formulae (17-18), \mathbf{r}_{\perp} is seen to remain unchanged in the general case.

What is of more fundamental significance is the observation that the spatial and temporal co-ordinates get mixed up in a Lorentz transformation, which tells us that the commonly perceived distinction between space and time is not an absolute one. In particular, the Newtonian concept of absolute time, independent of the observer, is fundamentally revised in the relativistic context. For instance, the concept of *simultaneity* of two events turns out to be a relative one, depending on the frame of reference of the observer. This is a simple consequence of the Lorentz transformation formula, as we see below.

17.2.5 Consequences of the Lorentz transformation formula

In exploring a number of simple consequences of the Lorentz transformation formula, we will consider the special case of two frames of reference S and S' wherein Cartesian co-ordinate systems are chosen as shown in fig. 17-1, in which case two of the four space-time co-ordinates (namely, y and z) of an event become redundant, since these do not get transformed in the change of the frame of reference. This is not any undue restriction since it only involves a special choice of co-ordinate axes in S and S', which

does not alter the physics of phenomena in any fundamental way. In any case, all the conclusions of physical relevance arrived at on the basis of the special case depicted in fig. 17-1 can be seen to remain valid under the more general co-ordinate transformation formulae given in sec. 17.2.4.

When we speak of the transformation of the space-time co-ordinates of events, we speak of the (Lorentz) transformation *formulae* (plural) while, in referring to the transformation of all the space-time co-ordinates considered as a single entity (say, in the form of a column) one, at times, speaks of the (Lorentz) transformation *formula* (singular) which, in general, is of the form (17-4a) with $L(V)$ as the transformation matrix.

Incidentally, *the velocity of light in free space (c) commands a very special position in the relativistic theory.* In the first place, it remains the same in all inertial frames of reference. This, indeed, is one central assumption on the basis of which the Lorentz transformation formula was arrived at. We will see later in connection with the *velocity transformation formulae* of the special theory of relativity (sec. 17.2.5.4) that a velocity v as measured in an inertial frame S , when considered in another inertial frame S' , gets transformed to a value v' that cannot exceed c (provided that v is less than c to start with).

Further, it takes an infinite amount of energy to increase the velocity of a particle of finite mass to the value c , which sets *the upper limit of all attainable velocities.* No particles moving with velocity greater than c have ever been found, which is consistent with the theory of relativity, and no consistent theoretical scheme has been proposed that accommodates particles faster than light as also slower ones.

Terms like 'velocity of light' or 'velocity of a particle' are often used to mean the speed rather than velocity in the sense of a vector. If a signal moves with speed c in any given direction in a frame S , it will be found to move with the same speed in any other frame S' , but possibly in a different direction. This change in the direction of propagation of

light, which is a consequence of the Lorentz transformation formula, is referred to as *aberration*.

17.2.5.1 Relativity of simultaneity

Consider two events with space-time co-ordinates (ct, x_1) and (ct, x_2) ($x_1 \neq x_2$) in S, where the equality of the time co-ordinate tells us that the events are simultaneous in this frame.

The temporal co-ordinates of the same two events in S' are, according to the co-ordinate transformation formulae

$$ct'_1 = \gamma(ct - \beta x_1), \quad ct'_2 = \gamma(ct - \beta x_2), \quad (17-19)$$

which differ from each other since, by assumption, $x_1 \neq x_2$ (the two events under consideration cease to be distinct if $x_1 = x_2$.)

In other words, a pair of events simultaneous in one frame are not perceived to be simultaneous by an observer in a different frame. This *relativity of simultaneity* is one indication that the concept of absolute time is a flawed one.

However, while the pre-relativistic concept of absolute time needs modification, there does remain a certain distinction between time and space, in the form of an absolute distinction between *time-like* and *space-like* separations, as we will see below (sec. 17.2.6.4). This is indicative of the fact that the *causal connection* between two events signifies a relation independent of the observer's frame.

Problem 17-4

Referring to formulae (17-19), consider two simultaneous events in frame S, separated by a distance l along the x-axis. What should be the velocity of the frame S' relative to S such that the time interval between the same two events in S' will be $\frac{l}{c}$. What will be spatial separation between the two events in S'?

Answer to Problem 17-4

HINT: By formulae (17-19), the time interval between the two events in S' is given by $t'_2 - t'_1 = -\frac{1}{c}\gamma\beta(x_2 - x_1) = -\gamma\beta\frac{l}{c}$. Hence the required velocity V of S' relative to S is given by $\gamma\beta = -1$, i.e., $V = \frac{c}{\sqrt{2}}$ along the *negative* direction of the x -axis. With $\beta = -\frac{1}{\sqrt{2}}$, the spatial separation in S' is seen to be $l' = x'_2 - x'_1 = \gamma l = \sqrt{2}l$.

17.2.5.2 Lorentz contraction

Suppose an observer in the frame S attempts to measure the length of a rod moving with velocity V , where the length of the rod and the direction of its velocity are both parallel to the x -axis. This means that the frame S' of a second observer, for whom the rod is at rest, is related to S as in fig. 17-1.

In measuring the 'length of the moving rod', the observer in S has to first *define* what the term means. A consistent and meaningful definition would correspond to the following procedure: the observer locates the two ends of the rod at the same instant of time, as recorded in her frame. This corresponds to two simultaneous events in S , one occurring at each end of the rod. Let the space-time co-ordinates of these two events be (ct, x_1) and (ct, x_2) respectively, where these correspond to the front end and the rear end of the rod (respectively, A and B, see fig. 17-3). She would then assign the value

$$l \equiv x_2 - x_1, \quad (17-20)$$

to the length of the moving rod.

The same two events would have space-time co-ordinates (ct'_1, x'_1) and (ct'_2, x'_2) in S' , where, by the Lorentz transformation formulae, x'_1, x'_2 are given by

$$x'_1 = \gamma(x_1 - \beta ct), \quad x'_2 = \gamma(x_2 - \beta ct). \quad (17-21)$$

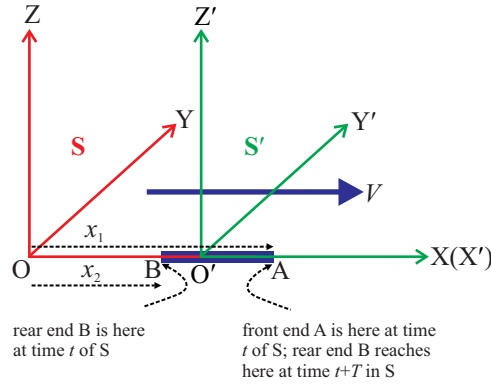


Figure 17-3: Measuring the length of a moving rod; the front end A of the rod crosses the point with co-ordinate x_1 in S at time t at which time the rear end B is located at x_2 ; the length of the moving rod is $l = x_1 - x_2$; the events with space-time co-ordinates (ct, x_1) and (ct, x_2) occur in S' at locations x'_1, x'_2 given by (17-21), in terms of which the *proper length* is given by (17-22); the rear end B of the rod reaches the point x_1 in S at time $t + T$ ($T = \frac{l}{V}$), which corresponds to an event occurring at location x'_1 in S' ; the proper length is then given by formula (17-25a).

In the frame S' , the two events would not be simultaneous as you can verify by working out the values of ct'_1, ct'_2 in accordance with the Lorentz formulae (check this out). *Regardless*, the observer in S' , would identify the difference

$$l_0 \equiv x'_1 - x'_2, \quad (17-22)$$

as the length of the rod since, in her frame, the rod is at rest and it does not matter whether she locates the ends of the rod at the same time or at different times (if she were to locate the rear end of the rod with the help of an event at time t'_1 , her result would still have the same value x'_2). This we term the *proper length* of the rod where the term ‘proper’ indicates that it is the length measured in the *rest frame* of the rod that is being referred to (which is why the symbol l_0 has been used in (17-22) instead of l' , which would have been more in keeping with the notational convention I have been using so far).

Making use of the transformation (17-21) in (17-20) and (17-22) one arrives at the relation

$$l = \gamma^{-1}l_0 = \sqrt{1 - \beta^2}l_0, \quad (17-23)$$

In other words, the length of the rod, as measured in the frame S (the one with respect to which the rod is moving with a velocity $V = c\beta$) turns out to be smaller than the proper length of the rod by a factor $\gamma = \sqrt{1 - \beta^2}$ (since c is the highest attainable velocity, the velocity V of the rod is necessarily less than c , and β is less than unity, i.e., $\gamma > 1$; only light signals in free space possess the velocity c).

This shortening of length as measured in a moving frame compared with the proper length is termed *Lorentz contraction* (or ‘length contraction’). Note that, in the above derivation, the length l_0 was assumed to be along the direction of the velocity of S’ relative to S (i.e., the direction of velocity of the moving rod). If the rod were assumed to be oriented in a direction perpendicular to the velocity, its length would come out to be l_0 in S, i.e., *there is no Lorentz contraction in the transverse direction*. This is a consequence of the fact that spatial co-ordinates measured along directions perpendicular to the relative velocity do not change in a Lorentz transformation, as seen from (17-2b).

There is another way that the length of a moving rod can be defined and measured in a given frame, that leads to the same length contraction formula as above. Consider the event, referred to above, occurring in S at time t corresponding to the front end A of the rod crossing the point x_1 (see fig. 17-3). Consider now a second event, corresponding to the rear end B of the rod crossing the *same* point x_1 of S. If the time of this event, as measured in S be, say $t + T$, then the observer in S would assign the length

$$l = VT = \beta cT, \quad (17-24)$$

to the moving rod, analogous to the way the length of a moving train is measured by noting the time interval between its front and rear ends moving past a fixed mark on the platform.

The times of occurrence (t' and $t' + T'$) of these two events, as observed in S’, can be worked out from the Lorentz formulae, but is not of relevance for our present purpose.

What is the location of the rear end of the rod in S' , given that its front end is at x'_1 (related to x_1 as in the first formula in (17-21)) at time t' ? This, evidently, is

$$x''_1 = x'_1 - l_0, \quad (17-25a)$$

regardless of the time interval between the two events under consideration because the rod, of proper length l_0 is at rest in S' . At the same time, one has

$$x''_1 = \gamma(x_1 - \beta c(t + T)). \quad (17-25b)$$

in accordance with the Lorentz formula, which thereby gives back the formula (17-23), i.e., the same relation between the proper length of the rod and its length, as measured in a moving frame (check this out; note that the velocity of the frame S with respect to the rest frame of the rod is $-V$, but the length contraction is independent of the sign of V). This, at the same time, indicates the consistency of the two ways, considered in this section, of assigning a length to the moving rod (each of which is further consistent with the fact that the length l tends to the proper length l_0 in the limit of $V \rightarrow 0$).

There sometimes arise apparent paradoxes from a wrong interpretation of the length contraction formula and other consequences of the Lorentz transformation which provides a relation between the space-time descriptions of events in two different frames. The spatial or temporal intervals between two events may get transformed in a change of the reference frame (though the *space-time separation* s occurring in (17-9) remains invariant). The paradoxes are resolved when one clearly identifies in physical terms the events involved in the transformation.

17.2.5.3 Time dilatation

Consider two events occurring at the same point x in the frame S , at time instants t_1, t_2 . The time interval

$$\tau = t_2 - t_1, \quad (17-26)$$

is referred to as the *proper time* between the two events since these two occur at the

same location in the frame S.

Considering now the frame S' that moves with velocity V relative to S (refer to fig. 17-1), the same two events occur at time instants

$$t'_1 = \gamma(t_1 - \beta cx), \quad t'_2 = \gamma(t_2 - \beta cx). \quad (17-27)$$

The time interval between the events in the moving frame S' is thus related to the proper time as

$$\tau' = t'_2 - t'_1 = \gamma(t_2 - t_1) = \gamma\tau. \quad (17-28)$$

In other words, the time interval gets *dilated* as one transforms from S to the moving frame S' (recall that γ is necessarily greater than unity, since $V < c$). The proper time, which is the time interval as measured in a frame in which the two events occur at the same location, is the *smallest* of the intervals measured in all possible inertial frames. This consequence of the Lorentz transformation formula is referred to as *time dilatation*.

Consider two inertial frames of reference S₁, S₂, from each of which two given space-time events are observed. Let the time interval between the events, as measured in the two frames, be τ_1 and τ_2 . Let now S₀ be the frame in which the same two events occur at the same spatial location, the time interval in this frame being τ_0 . The latter is then the proper time interval between the two events. If V₁ and V₂ be the velocities of S₁, S₂ relative to S₀ then, according to (17-28), one has

$$\gamma_1\tau_1 = \gamma_2\tau_2 = \tau_0, \quad (17-29a)$$

where

$$\gamma_i = \frac{1}{\sqrt{1 - \frac{V_i^2}{c^2}}} \quad (i = 1, 2). \quad (17-29b)$$

In other words, given any two space-time events, the quantity $\gamma\tau$, referred to any arbitrarily chosen inertial frame, is an *invariant* one, i.e., remains unchanged when considered in any other inertial frame since, for each such frame, it is nothing but the proper time interval between the two events. Here γ is defined as in (17-29b) (i.e., $\gamma = \frac{1}{\sqrt{1 - \frac{V^2}{c^2}}}$),

where V is the velocity of the frame under consideration with respect to the frame in which the two events occur at the same spatial location).

Strictly speaking, the above considerations apply in the case of two events for which the space-time separation is a *time-like* one (you will find the idea of *time-like* and *space-like* separations between events introduced in sec. 17.2.6.4), because it is only for such a separation that the concept of a proper time is meaningful. However, the relations (17-29a) remain valid in a formal sense even for a pair of events for which the space-time separation is space-like since, in that case, both γ and τ reduce to *imaginary* quantities. This is because of the fact that the frame S_0 in which the events occur at the same spatial location is now no more than a notional one since the velocity of that frame with respect to either of the frames S_1, S_2 under consideration turns out to be larger than c . This is beyond the limit, set by the theory of relativity, of physically possible velocities.

Problem 17-5

Referring to fig. 17-1, let the time interval between two events, as observed in the inertial frame S , be τ , and the spatial separation between the two be l , where it is assumed that both the events occur on the x -axis of S . Find the velocity V of a second frame S' for which the two events occur at the same point, and the proper time interval τ_0 , expressing both in terms of l and τ , and obtain the condition for the second frame S' to exist.

Answer to Problem 17-5

HINT: Let the space-time co-ordinates of the two events in S be (ct, x) and $(ct + c\tau, x + l)$. Let the space-time co-ordinates of the same two events in S' be (ct', x') and $(ct' + c\tau_0, x')$ where the spatial co-ordinate of the second event is the same as that of the first, in which case the time interval is the proper time interval between the two events. Then, invoking the Lorentz transformation formulae for the space- and time intervals, we obtain, $c\tau_0 = \gamma(c\tau - \beta l)$, and $0 = \gamma(l - \beta c\tau)$. These relations give $\tau = \frac{l}{V}$, $\tau_0 = \frac{\tau}{\gamma}$ (as expected), i.e., $\tau_0 = \tau \sqrt{1 - \frac{l^2}{c^2\tau^2}}$. The condition for the second frame S' to exist is that its speed $|V|$ is to be less than c , i.e., $l^2 < c^2\tau^2$, which means that the

squared space-time interval $s^2 = c^2\tau^2 - l^2$ is to be positive. As we see below in sec. 17.2.6.4, this is expressed by saying that the space-time separation between the two events has to be a *time-like* one.

A paradox owing its origin to the above time dilatation formula - a much discussed one in the literature - is the so-called *twin paradox*. One of two identical twins is assumed to reside in an inertial frame S while the second twin starts out, at time $t' = t = 0$ (say) with velocity V in a space ship, to which is attached a frame S' . The latter travels in a straight line (there are variants of the problem formulation, of which I choose one) for a proper time $\frac{T}{2}$, after which it instantaneously reverses its direction of motion, and again meets the first member of the twin after a second interval of $\frac{T}{2}$, i.e., after the total time interval T , as measured in S' . This, however, corresponds to an interval γT in S. Thus, the first of the twins ages more, by a factor of γ , as compared to the second member. But, from the point of view of the latter, it is the *first member* that has made the space travel and so, it is the first member who should age less. This apparent lack of symmetry between the two frames constitutes the paradox.

The paradox is resolved entirely within the confines of the special theory of relativity, without any special reference to the acceleration of S' relative to S, though it is the acceleration that distinguishes the two frames.

You will find a careful analysis of the twin paradox in [13], where a number of common misconceptions regarding the resolution of the paradox are cleared up.

17.2.5.4 Velocity transformation

Consider the motion of a particle as observed in an inertial frame S in which its position co-ordinates at time t , with respect to any chosen Cartesian co-ordinate system are (x, y, z) . Let the co-ordinates at time $t + \delta t$ be $(x + \delta x, y + \delta y, z + \delta z)$ where the symbol δ denotes here, as elsewhere in this chapter, a small increment which is assumed to tend to zero when the ratio of two such increments is involved, resulting in the relevant *derivative*. Thus, the components of the instantaneous velocity \mathbf{v} of the particle in the

frame S are

$$v_x = \frac{\delta x}{\delta t}, \quad v_y = \frac{\delta y}{\delta t}, \quad v_z = \frac{\delta z}{\delta t}. \quad (17-30)$$

Let us now consider the motion of the same particle, as observed in a second frame S' where the latter is assumed to be related to S as in fig. 17-1 for the sake of simplicity. Then the events with space-time co-ordinates (ct, x, y, z) and $(c(t+\delta t), x+\delta x, y+\delta y, z+\delta z)$ in S, will have space-time co-ordinates (ct', x', y', z') and $(c(t'+\delta t'), x'+\delta x', y'+\delta y', z'+\delta z')$ in S', where the primed quantities are related to the unprimed ones as in (17-4a), (17-4b). Making use of the fact that these transformations are linear, one obtains

$$c\delta t' = \gamma(c\delta t - \beta\delta x), \quad \delta x' = \gamma(\delta x - \beta c\delta t), \quad \delta y' = \delta y, \quad \delta z' = \delta z, \quad (17-31)$$

(check this out).

The components of the instantaneous velocity of the particle, as measured in S', are defined by formulae analogous to those in (17-30), with the primed quantities replacing the corresponding unprimed ones. Using the transformation relations (17-31) in these, one arrives at

$$v_x' = \frac{\delta x'}{\delta t'} = \frac{v_x - V}{1 - \frac{Vv_x}{c^2}}, \quad v_y' = \frac{\delta y'}{\delta t'} = \frac{v_y}{\gamma(1 - \frac{Vv_x}{c^2})}, \quad v_z' = \frac{\delta z'}{\delta t'} = \frac{v_z}{\gamma(1 - \frac{Vv_x}{c^2})}, \quad (17-32)$$

(check these out). These give the *velocity transformation formulae* resulting from the Lorentz transformation from S to S'.

Note that these reduce to the Galilean velocity transformation formula (3-41) in the limit of small velocities, i.e., for small values of $\frac{V}{c}$, $\frac{v_x}{c}$, $\frac{v_y}{c}$, $\frac{v_z}{c}$, such that terms of degree higher than one in these velocity ratios can be ignored (check this out).

The conditions of smallness of $\frac{v_y}{c}$, $\frac{v_z}{c}$ are not required in establishing the limiting transition from (17-32) to the Galilean transformation formula (3-41). However, these are found to be relevant in the more general case when the co-ordinate systems in S and S' have an arbitrarily chosen orientation relative to each other, and the relative velocity is along an arbitrarily chosen direction (refer to (17-33) below).

The velocity transformation formulae can also be derived in the more general case when the co-ordinate systems in S and S' have an arbitrarily chosen orientation relative to each other, and the relative velocity is along an arbitrarily chosen direction, by making use of the Lorentz transformation formulae (17-4a), (17-17) (refer to sec. 17.2.4). A compact expression for such a general velocity transformation formula is obtained from (17-18) in the form

$$v_{\parallel}' = \frac{v_{\parallel} - V}{1 - \frac{Vv_{\parallel}}{c^2}}, \quad \mathbf{v}_{\perp}' = \frac{\mathbf{v}_{\perp}}{\gamma(1 - \frac{Vv_{\parallel}}{c^2})}, \quad (17-33)$$

where the suffixes ' \parallel ' and ' \perp ' denote, respectively, components parallel and perpendicular to the direction of the relative velocity (check the above formulae out).

It may be noted that, in the transformation from S to S', the transverse spatial co-ordinates remain unaltered, but the transverse components of the *velocity* of a moving particle get transformed. This is due to the fact that the velocity components involve a time derivative, and time intervals get altered in the transformation.

Problem 17-6

Of two uniformly moving particles A and B, the former is located at $x = 0$ in the frame S, and the latter is located at $x' = a(> 0)$ in S', both at $t = t' = 0$, when the velocity of A in S is v and that of B in S' is v' (all velocities parallel to the x-axis in S, S', being defined with reference to the positive direction of the latter, i.e., a motion along the negative direction of the x-axis implies a negative value of the corresponding velocity). Given that the velocity of S' with respect to S is V , find the time t (in S) at which the two particles collide, and the minimum value of v for which a collision takes place.

Answer to Problem 17-6

HINT: By the Lorentz transformation formulae, the location of the particle B in the frame S at time $t = t' = 0$ is at $x = \gamma a$, while its velocity in S is, by the velocity transformation rule $v'' = \frac{v' + V}{1 + \frac{v'V}{c^2}}$ (note that the velocity of S with respect to S' is $-V$). The time at which the collision takes place is thus $t = \frac{x}{v - v''} = \frac{\gamma a}{v - \frac{v' + V}{1 + \frac{v'V}{c^2}}}$. The condition for the collision to occur is $v > \frac{v' + V}{1 + \frac{v'V}{c^2}}$.

17.2.5.5 Relativistic aberration

Consider the propagation of light, as observed in two inertial frames S , S' where, for the sake of simplicity, S' is assumed to be related to S as in fig. 17-1. We choose co-ordinate axes in S and S' such that the direction of propagation is in the x - y plane of either system. Let the direction of propagation make an angle θ with the x -axis in S as shown in fig. 17-4, in which the direction is indicated by the arrowhead marked 'A'. Since the speed of light is c in all inertial frames, the components of the velocity vector describing the propagation of light in S are given by

$$v_x = c \cos \theta, \quad v_y = c \sin \theta, \quad v_z = 0, \quad (17-34)$$

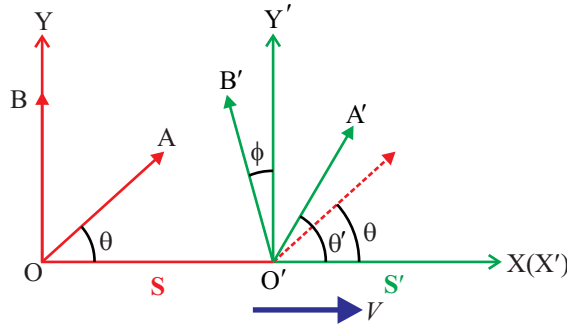


Figure 17-4: Illustrating the phenomenon of relativistic aberration; the frame S' is related to S as in fig. 17-1; the propagation of a plane wave is considered, along the direction marked 'A', as observed in S , the angle between this direction and the x -axis being θ ; to an observer in S' , the propagation appears to take place along the direction marked 'A'', which makes an angle θ' with the x -axis; θ' differs from θ , being related to the latter as in eq. (17-37); the difference in the directions of propagation in the two frames is termed 'aberration'; in the particular case of the propagation in S being along the y -axis (direction 'B', $\theta = \frac{\pi}{2}$), that in S' will be along the direction 'B'' ($\theta' = \frac{\pi}{2} + \phi$), where the angle ϕ is given by (17-38), (17-39).

On applying the velocity transformation formulae (17-32), one obtains the velocity components in S' as

$$v'_x = \frac{c \cos \theta - V}{1 - \frac{V \cos \theta}{c}}, \quad v'_y = \frac{c \sin \theta}{\gamma(1 - \frac{V \cos \theta}{c})}, \quad v'_z = 0, \quad (17-35)$$

where γ is given by (17-2c).

One corollary of these relations is that the speed of light in S' works out to c , as it

should, by the relativistic principle of invariance of the speed of light:

$$v_x'^2 + v_y'^2 + v_z'^2 = c^2, \quad (17-36)$$

(check this out), while the same formulae tell us that the angle θ' made by the direction of propagation with the x-axis in the frame S' (see figure, direction marked A') is given by

$$\tan \theta' = \frac{v_y'}{v_x'} = \frac{c \sin \theta}{\gamma(c \cos \theta - V)}. \quad (17-37)$$

This shows that the direction of propagation of light gets changed in a Lorentz transformation even though the speed of light remains invariant. This phenomenon is referred to as *relativistic aberration*. As a simple instance, consider the particular case $\theta = \frac{\pi}{2}$, which means that the propagation takes place along the y-axis in the frame S (see figure, direction marked 'B'). In the frame S' , on the other hand, the direction of propagation is tilted away from the y-axis in S' (marked 'B'' in figure) by an angle ϕ (i.e., the tilt with respect to the x-axis is $\frac{\pi}{2} + \phi$). The angle of tilt ϕ is given by

$$\tan \phi = \frac{\gamma V}{c} \quad (\theta = \frac{\pi}{2}), \quad (17-38)$$

(check this out). This differs (by terms of degree two and higher in $\frac{V}{c}$) from the non-relativistic formula as obtained by the application of the Galilean velocity transformation rule

$$(\text{non - relativistic}) \quad \tan \phi = \frac{V}{c} \quad (\theta = \frac{\pi}{2}), \quad (17-39)$$

(check this out) where it is to be noted that the Galilean formula is inconsistent with the invariance of the speed of light in a change of the frame of reference.

17.2.6 Space-time diagrams and world lines

17.2.6.1 Representation of events and world lines

Given a co-ordinate system (which we assume to be a Cartesian one) in an inertial frame S and appropriate time-keeping clocks in it, any event can be specified in terms of four space-time co-ordinates, say, ct, x, y, z , where the time co-ordinate is scaled by a factor c

for the sake of convenience and of uniformity of dimension.

Let us consider, for the sake of simplicity, only those events for which the co-ordinates y, z have specified values, say, $y = 0, z = 0$. For instance, we can focus attention on a particle moving along the x -axis of the co-ordinate system chosen. The fact of the particle being at a location, say, x at time t then constitutes an event that can be depicted with just two space-time co-ordinates (ct, x) in a two dimensional *space-time diagram*, as shown in fig. 17-5. Of the two axes shown, the horizontal axis corresponds to the spatial co-ordinate x and the vertical axis to the scaled time co-ordinate ct .

It is a matter of notational convention (not followed universally) that, while indicating the co-ordinates of a point in the space-time diagram, the co-ordinate measured along the vertical axis is mentioned first while that measured along the horizontal axis comes next. This differs from the convention commonly followed in co-ordinate geometry, but is not likely to cause confusion once one gets to be aware as to what is what.

In this figure, E_1 and E_2 represent two events with space-time co-ordinates (ct_1, x_1) and (ct_2, x_2) in S . The straight line A represents a continuous succession of events, and corresponds to the uniform motion of a particle that crosses the origin $x = 0$ at time $t = 0$, the angle ϕ shown in the figure being determined by the relation

$$\tan \phi = \frac{c}{v}, \quad (17-40)$$

for the slope of the line, where v stands for the velocity of the particle (check this out; v is assumed positive in the figure). The line A is referred to as the *world line* depicting the motion of the particle. The line B depicts another world line, corresponding to a uniform motion with a velocity larger than v , while C, C' , having slopes ± 1 , represent world lines for motions with speed c (the maximum attainable speed), respectively along the positive and the negative directions of the x -axis. In the same figure, D depicts a world line for a motion with *non-uniform* velocity where, as in the case of A, B, C , and C' , we assume that the origin ($x = 0$) is crossed at time $t = 0$.

Since the speed of a particle is necessarily less than c at all times, the world lines corresponding to all possible motions (satisfying the condition $x = 0$ at $t = 0$) are confined within the region between the lines C, C' , made up of the regions marked F and P in the figure (reason this out).

The condition $x = 0$ at $t = 0$ is not a restrictive one since, for any world line such as D satisfying this condition, one can obtain another world line for which $x = x_0$ at $t = 0$, simply by shifting it along the horizontal axis by a distance x_0 . A set of world lines obtained by such a shift is shown in the figure by dotted lines. The regions F and P get translated to F' and P' , bounded by lines obtained from C, C' by the horizontal shift through x_0 .

Instead of talking in terms of motions of particles, one may talk of world lines relating to *signals*, since a moving particle can be used as a signal between an initial event and a final event. The lines C, C' then correspond to *light signals* along the positive and negative directions of the x -axis, where these may be interpreted as signals carried by *photons*.

These two world lines (C and C' in the figure) divide up the entire space-time diagram into three *regions*, two of which are the regions F and P mentioned above, while the third region is the complementary one, lying to the left and right of F and P (marked C_1, C_2 respectively).

These regions, and the lines C, C' making up the boundaries of these regions, are *invariant ones* in a certain sense. This we will see in sec. 17.2.6.3 below, but only after having a look at *transformations* in the space-time diagram.

17.2.6.2 The space-time diagram and Lorentz transformations

The points E_1 and E_2 in fig. 17-5 denote two different events, with distinct sets of co-ordinates, as measured in the frame S . For an arbitrary choice of E_1, E_2 , the pairs of co-ordinates (ct_1, x_1) and (ct_2, x_2) need not be related to each other. But the same space-

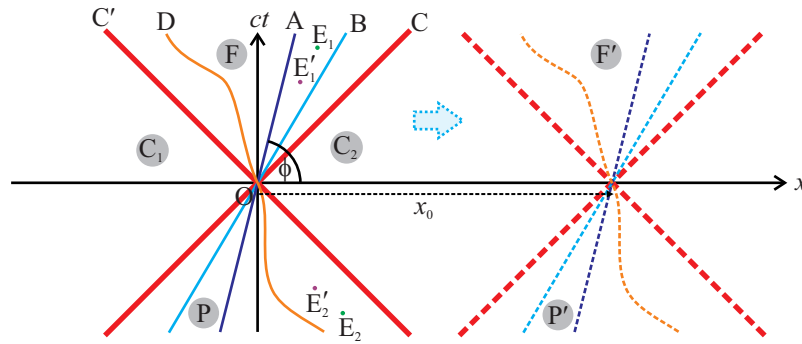


Figure 17-5: Illustrating the idea of a space-time diagram and of events and world lines in it; we consider, for the sake of reference, a Cartesian co-ordinate system in an inertial frame S and (again for the sake of simplicity) events for which the y - and the z - co-ordinates are redundant, such as those relating to the motion of a particle along the x -axis; the diagram consists of two axes perpendicular to each other, corresponding respectively to the space-time co-ordinates x and ct ; points in the diagram such as E_1 and E_2 represent events, as observed in S ; a continuous succession of events corresponds to a world line; A, B, C, C' , and D are a few world lines, among which C, C' , with slopes ± 1 , correspond to light signals passing through $x = 0$ at $t = 0$; world lines passing through $x = x_0$ at $t = 0$ are obtained by a translation along the horizontal axis; while all these events and world lines refer to a single frame S , one can also consider a Lorentz transformation from S to S' , under which one gets a transformed set of points and world lines, depicted in the same space-time diagram (e.g., E'_1, E'_2 from E_1, E_2); F and P are invariant regions that get transformed into themselves; the complementary region, indicated by symbols C_1, C_2 on the left and right, is also an invariant one.

time diagram may be used to represent events and world lines as observed in a second inertial frame S' as well, with the position co-ordinate, as observed in S' , measured along the horizontal axis and the scaled time co-ordinate in S' along the vertical axis.

In this alternative view, the points E'_1, E'_2 in fig. 17-5 denote the transformed points corresponding to E_1, E_2 respectively, which means that their space-time co-ordinates are related to those of E_1, E_2 by Lorentz transformation formulae as in sec. 17.2.2.

You need to be clear in your mind as to what is involved here. A space-time diagram is, by definition, a representation, by means of space-time co-ordinates, of events and world lines as observed in a given frame of reference such as S . When one speaks of a representation, in the *same* diagram, by means of space-time co-ordinates as observed in a *second* frame such as S' , one is playing a tricky game. Take the case of the point E_1 as an example. For the sake of concreteness, let the space-time co-ordinates of this event, as observed in S' correspond to the point E'_1 when plotted in the same

space-time diagram as the one shown in the figure. Thus, the points E_1, E_2 denote distinct space-time events, while E_1, E'_1 (or E_2, E'_2) denote one and the same event, now looked at from two different frames. On the other hand, from the point of view of an observer in S , without regard to S' , E_1 and E'_1 may equally well be interpreted as points representing two distinct events. The important thing to remember is that events and world lines are the basic physical entities, and the space-time co-ordinates in any chosen frame are book-keeping entries, indicated in a space-time diagram. If the same space-time diagram is used for two different frames, quantities belonging to two different sets (one for each frame) are being indicated along the horizontal, as also along the vertical, axis.

17.2.6.3 The invariant regions

The regions F and P, taken together, are described by the formula

$$c^2t^2 > x^2, \quad (17-41)$$

for any point with co-ordinates (ct, x) lying in it, where these co-ordinates refer to an arbitrarily chosen event observed in the frame S . Now consider a Lorentz transformation to a second frame S' , by which one obtains a transformed point (corresponding, however, to the same *event*) with co-ordinates (ct', x') , where the co-ordinates pertaining to S' are now being depicted in the same space-time diagram as the one meant for S .

Now, in accordance with the invariance of the squared space-time interval in a Lorentz transformation (see sec. 17.2.3), one has

$$c^2t^2 - x^2 = c^2t'^2 - x'^2, \quad (17-42)$$

since, in the present context (where S and S' are related as in fig. 17-1), the co-ordinates y, z do not change under the transformation. This means that the inequality (17-41) is satisfied for the primed co-ordinates as well, and so the point with the transformed co-ordinates lies in the composite region made up of F and P. A similar argument tells us that an initial point chosen to lie in the complementary region (marked C_1, C_2 on the left

and right in fig. 17-5) is also an invariant one. Finally, the lines C, C' representing light signals moving in the positive and negative directions of the x-axis are also invariant under a Lorentz transformation.

What is more, the regions F and P are *separately* invariant under Lorentz transformations. For instance, the co-ordinates of a point lying in the region F satisfy (17-41) and, in addition, the inequality

$$t > 0, \quad (17-43a)$$

i.e., F includes events lying in the *future* as compared with the event at $ct = 0, x = 0$. Together, these two inequalities imply $ct > \frac{V}{c}x$ (reason this out; V can be negative if the relative velocity is along the negative direction of the x-axis), i.e., by the transformation equation (17-2a),

$$t' > 0. \quad (17-43b)$$

This means that the transformed point (with co-ordinates as observed in S') also lies in F. An analogous reasoning leads to the same conclusion for the region P (which includes *past* events as compared with the event at O).

The lines C, C' are said to constitute the *light cone* since these are the world lines of light signals passing through the origin. Within the light cone, the regions F and P, each of which is invariant under Lorentz transformations, are made up of future and past events relative to O.

Thus, no Lorentz transformation can take an event-point from P to F or from F to P: *the future and the past are absolute concepts, independent of the frame of reference*. This is related to the concept of *causality*. The event at O ($ct = 0, x = 0$) can be made to act as the cause of an event such as E_1 , which appears as the effect, since a signal originating at O can connect to an event at a time in the future, but not to one (such as E_2) in the past, since the principle of causality tells us that the effect cannot occur at a time earlier than the cause (however, the event E_2 can be the cause for O, which then appears as the effect).

17.2.6.4 Time-like and space-like separations

Something more can be said of the invariant regions F and P. For any point E_1 at (ct_1, x_1) in F, for instance, one can find a Lorentz transformation such that the transformed point E'_1 (see fig. 17-6) has its position co-ordinates $x'_1 = 0$, which means that a frame can be chosen in which the event under consideration (remember that E_1 and E'_1 relate to the *same* event, as observed in the two frames S and S') occur at the *same* spatial location as the reference event at O, only at a later time (refer to problem 17-5). The time interval $t'_1 (= \gamma(t_1 - \frac{Vx_1}{c^2}))$, i.e., the interval between O and E'_1 is the *proper* time interval between the two events (refer back to sec. 17.2.5.3), and gives a measure of the invariant space-time separation between the events with space-time co-ordinates $(0, 0)$ and (ct_1, x_1) in S. Such a space-time separation is termed a *time-like* one.

Analogous statements hold for an event corresponding to a point $E_2 (ct_2, x_2)$ lying in P. Once again, a Lorentz transformation can be found in which the event occurs at $x'_2 = 0$, i.e., at the same spatial location as the reference event at O $(0, 0)$. The time interval between the events in the new frame then gives the proper time interval, and determines the invariant space-time separation between the events, which is termed a time-like one.

The story is quite different for the invariant region marked C_1, C_2 to the left and right in figures 17-5, 17-6. In the first place, C_1 and C_2 are *not* separately invariant and, together, make up one single invariant region. This means that a point Q in C_1 (resp. C_2) can be transformed to one in C_2 (resp. C_1) by an appropriate Lorentz transformation (reason this out). What is more, since the velocity of any signal necessarily satisfies the inequality $|v| < c$, the events at O and at Q cannot be causally connected. The space-time separation between the events at O and Q is imaginary (the squared space-time separation is negative) and is said to be a *space-like* one.

Finally, an event corresponding to a point lying on the light cone, made up of the lines C, C', can be causally connected with the event at O only by means of a light signal, the space-time separation between the two events being zero. This is referred to as a

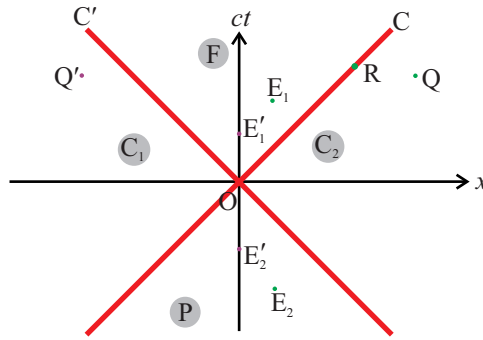


Figure 17-6: Explaining the idea of time-like and space-like separations; E_1 and E_2 are points in the space-time diagram with co-ordinates (ct_1, x_1) and (ct_2, x_2) , lying in the invariant regions F and P respectively; one can find a Lorentz transformation by which E_1 is transformed to E'_1 , the latter having co-ordinates $(ct'_1, 0)$, where ct'_1 is the space-time separation between the event under consideration and that corresponding to the origin O; since these two events now occur at the same location, the interval between the two is said to be a time-like one; analogous statements hold for the event represented by E_2 , which can be transformed to a point E'_2 lying on the time axis; the region marked C_1 , C_2 to the left and right of the light cone is a single invariant region, and a point Q lying in any one of the two sub-regions can be transformed to one (Q') lying in the other; the space-time separation between O and Q is said to be a space-like one; in the case of an event represented by a point R lying on the light cone (made up of lines C, C'), its space-time separation from the event represented by the origin O is zero, and is said to a light-like one.

light-like separation.

17.2.7 Space-time geometry

17.2.7.1 Geometry in '1+1' dimensions

Fig. 17-7 depicts a plane with a pair of axes indicating the spatial co-ordinates (x, y) of any arbitrarily chosen point, which I have included here in order to make a comparison with the space-time diagram of fig. 17-5 where the latter has the appearance of a two dimensional space. However, the latter indicates the spatial location of a point (x) in one dimension and the time co-ordinate (ct) in the second (recall that fig. 17-5 pertains to the special situation where two of the three spatial dimensions are redundant). This contrast between the two diagrams is highlighted by saying that the x - y plane of fig. 17-7 is a 2-dimensional one while the x - ct plane of fig. 17-5 is (1+1)-dimensional.

The (1+1)-dimensional plane looks quite like the the 2-dimensional one, but the two differ in an essential respect. Recall that the geometry of the two dimensional plane, i.e., the relations between points, lines, and figures drawn in it depend fundamentally

on the *metric* characterizing it, namely, the *distance* between two neighboring points. Considering two points (x, y) and $(x + \delta x, y + \delta y)$, the squared distance $(\delta s)^2$ is given by the formula

$$(\delta s)^2 = (\delta x)^2 + (\delta y)^2. \quad (17-44)$$

This, essentially, follows from the Pythagoras theorem and from the fact that (x, y) are Cartesian co-ordinates in the plane.

One can make a different choice of co-ordinates, such as the plane polar co-ordinates (r, θ) , in terms of which the squared distance between two neighboring points appears as

$$(\delta s^2) = (\delta r)^2 + r^2(\delta \theta)^2. \quad (17-45)$$

More generally, for an arbitrary choice of co-ordinates (u_1, u_2) , the squared distance is a quadratic expression of the form

$$(\delta s)^2 = A(u_1, u_2)(\delta u_1)^2 + B(u_1, u_2)\delta u_1\delta u_2 + C(u_1, u_2)(\delta u_2)^2, \quad (17-46)$$

which involves the scale factors A, B, C , each of which depends, in general, on u_1, u_2 , and where there occurs the ‘cross term’ involving the product $\delta u_1\delta u_2$. The good thing about the 2-dimensional plane is that there exists a pair of co-ordinates (namely, the Cartesian ones: $u_1 = x, u_2 = y$) for which the cross term is absent and the scale factors are constants, independent of the co-ordinates.

What is of fundamental importance about the squared distance is that it is an *invariant* quantity under a set of co-ordinate transformations, namely, rotations of the co-ordinate axes. Fig. 17-7 shows a pair of rotated co-ordinate axes, for which the new co-ordinates are x', y' . In terms of these transformed co-ordinates, the squared distance is of the same form as in (17-44), telling us that the quadratic expression on the right hand side of the equation is invariant under the rotation:

$$(\delta s^2) = (\delta x)^2 + (\delta y)^2 = (\delta x')^2 + (\delta y')^2. \quad (17-47)$$

Now consider the (1+1)-dimensional space-time diagram of fig. 17-5, in which the lo-

cation of point is indicated by the co-ordinates (ct, x) . The geometry in this diagram is principally determined by the squared space-time separation of formula (17-6a) which, for a pair of neighboring points, assumes the form

$$\delta s^2 = (\delta(ct))^2 - (\delta x)^2. \quad (17-48)$$

This is similar to formula (17-44) in that the scale factors are independent of the co-ordinates - a characteristic feature of the space under consideration that we express by saying that it is a *flat* one. In the case of a flat space it is possible to find a set of co-ordinates for which the cross terms do not appear in the expression for δs^2 .

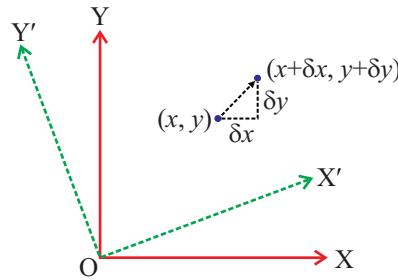


Figure 17-7: Depicting the 2-dimensional plane defined by the axes OX, OY of a Cartesian co-ordinate system, where the co-ordinates x, y indicate the spatial location of a point; the geometry in this plane differs essentially from that in the (1+1)-dimensional space-time diagram of fig. 17-5; the geometry is determined in either case by the metric, i.e., the squared separation (δs^2) between two neighbouring points $((x, y)$ and $(x + \delta x, y + \delta y)$ in the present figure); this is given by the positive definite expression of (17-44) for the 2-dimensional Cartesian plane, and by the indefinite expression of (17-48) for the (1+1)-dimensional space-time diagram; a second pair of axes (OX', OY'), obtained by applying a rotation to OX, OY is shown; the expression for the squared separation retains its form under the rotation, as expressed in (17-47); in the (1+1)-dimensional case, the squared separation between two points retains its form under a Lorentz transformation, which one can interpret as a rotation of the co-ordinate axes by an *imaginary* angle.

While the expression for the squared separation in the case of the 2-dimensional Cartesian plane remains invariant under a rotation of the co-ordinate axes as indicated above, in the case of the (1+1)-dimensional space, the squared space-time separation remains invariant under a Lorentz transformation, where the latter can be formally interpreted as a rotation of the two co-ordinate axes by an *imaginary* angle.

On the other hand, the expression (17-48) differs fundamentally from (17-44) in that the latter is a *positive definite* quadratic expression while the former is *not*. In other words, the squared space-time separation can have positive, negative, or zero value (corresponding to, respectively, time-like, space-like or light-like separations), while the squared distance in the 2-dimensional plane is always positive (it has the value zero only in the trivial case of δx and δy being both zero).

17.2.7.2 The (1+3)-dimensional space-time geometry

More generally, an event, as observed in an inertial frame S, is represented in a space-time diagram that can be described as a (1+3)-dimensional space, with three dimensions for the spatial location of the event and one dimension for its time of occurrence. The squared space-time separation between two neighboring events with space-time coordinates (ct, x, y, z) and $(c(t + \delta t), x + \delta x, y + \delta y, z + \delta z)$ is a quadratic expression of the form

$$\delta s^2 = (\delta(ct))^2 - (\delta x)^2 - (\delta y)^2 - (\delta z)^2. \quad (17-49)$$

This corresponds to a flat space (with scale factors $1, -1, -1, -1$, independent of the space-time co-ordinates) and is an *indefinite* quadratic expression that makes the geometry of the (1 + 3)-dimensional space of all possible space-time events (in the absence of gravitation) one of a very special character. The geometry in this space is referred to as the *Minkowski* geometry. The simplest co-ordinate system in terms of which one can describe geometrical relations in this ‘Minkowski space’ are the scaled time co-ordinate (ct) and the Cartesian space co-ordinates (x, y, z) , as measured in an inertial frame. While other sets of co-ordinates can, in principle, be used, making the quadratic form for δs^2 look more complex, the space-time co-ordinates pertaining to an inertial frame are the simplest ones.

The *light cone* in the Minkowski space is a three-dimensional surface given by the formula

$$c^2 t^2 - x^2 - y^2 - z^2 = 0, \quad (17-50)$$

and divides the Minkowski space into invariant regions as in the case of the (1+1)-

dimensional space-time diagram. Fig. 17-8 depicts the light cone in the relatively simpler case of a $(1 + 2)$ -dimensional Minkowski space, where only those space-time events are considered for which the co-ordinate z has a fixed value, say, $z = 0$ (this condition is invariant under those Lorentz transformations for which the relative velocity vector lies in the x - y plane).

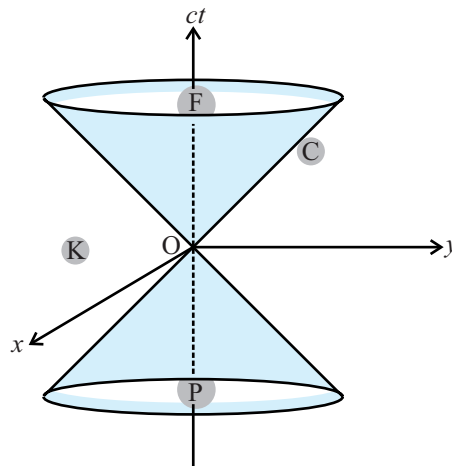


Figure 17-8: Depicting the light cone and the invariant regions for a $(1+2)$ -dimensional Minkowski space, in which one dimension is for the scaled time co-ordinate and the other two are for spatial co-ordinates; the third spatial co-ordinate is not considered for the sake of simplicity; the light cone C is an invariant conical surface, while the region interior to it is made up of two regions, each separately invariant under Lorentz transformations - the 'future' (F) and the 'past' (P); the complementary region K is also an invariant region which, unlike the $(1+1)$ case, is a connected one; an event within F or P has a time-like separation from O , while one in K has a space-like separation; analogous statements hold for the $(1+3)$ -dimensional Minkowski space of all possible space-time events.

As in the $(1 + 1)$ -dimensional case, the 'future' (F) and the 'past' (P) are two disjoint invariant regions in the interior of the light cone, for each of which the space-time interval between an arbitrarily chosen point in the region and the origin is a time-like one ($\delta s^2 > 0$). In the invariant region complementary to these two, lying outside the light cone and marked K in the figure, the squared space-time separation between any point and the origin is a space-like one ($\delta s^2 < 0$), and no event in this region can be connected causally to the event at the origin O . However, unlike the $(1 + 1)$ -dimensional case where the region of space-like separations is made up of two disjoint sub-regions (C_1 , C_2 in fig. 17-5), this region in the $(1 + 2)$ -dimensional Minkowski space is a *connected* one

in which any two points can be connected by a continuous curve lying wholly in this region. All these features of the $(1+2)$ -dimensional Minkowski space remain true for the $(1+3)$ -dimensional space as well.

17.2.8 Physical quantities as four-vectors

17.2.8.1 Vectors: the basic idea

In chapter 2, I gave you the basic idea underlying vectors, where I mentioned that the fundamental thing about vectors is their *linearity*. A vector is an element of a set (a linear vector space) in which the operation of vector *addition* is defined, along with the operation of *multiplication with a scalar* (vector addition, taken in conjunction with multiplication with scalars, leads to the more general operation of *linear combination* of vectors). A vector space always comes associated with a set (technically, a *field*) of *scalars*. Commonly, the scalars happen to be real numbers, and are not mentioned separately in the definition of a vector space.

You are by now familiar with one dimensional, two dimensional, and three dimensional vectors, that can be identified with directed line segments (i.e., *arrows*) along a line, in a plane, and in space respectively.

The term ‘space’ at times causes confusion. In mathematics and physics, a space is often used to denote a set of some particular kind (e.g., vector space) while another common use of the term is in the sense of the physical space in which we all reside. Let us, for the time being, use the term ‘position space’ while referring to this physical three dimensional space, which contains ‘position planes’, ‘position lines’, and ‘position points’. In expressions like $(1+2)$ -dimensional space or $(1+3)$ -dimensional space, the term ‘space’ is used in the more general mathematical sense, where a ‘point’ is determined by a time instant t , and a position point in a position plane (such as the x-y plane in a Cartesian co-ordinate system) or in the position space, as the case may be. In an expression such as ‘space-time diagram’, on the other hand, the term ‘space’ is used in the sense of our familiar position space. It sometimes needs a bit of care to grasp the intended meaning in a given context. There is, however, substantial overlap

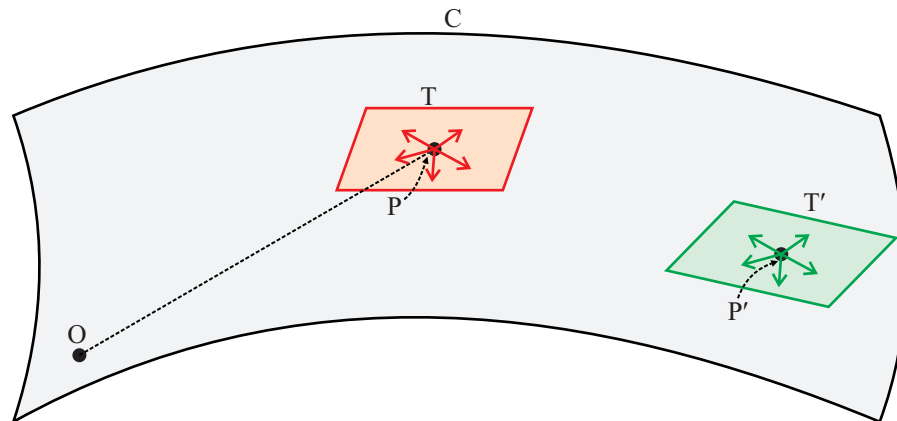


Figure 17-9: Illustrating the idea of a curved space and that of the tangent space at a point; C is a curved surface in three dimensional space, the latter being a flat one; the arrow directed from a chosen origin O up to the point P in C has no relevance in the context of C; however, imagining little arrows drawn at P, each along the tangent to C in some arbitrarily chosen direction, one arrives at the tangent space T at P, containing the set of vectors based at P; the tangent space T' at some other point P' is similarly made up of vectors based at P'.

between the two usages

Referring to arrows (directed line segments) in position space, one commonly associates a vector with a point in this space, called the *radius vector* for that point, which in reality is the arrow extending from the origin up to the point in question. However, the location of the tail (or of the tip) of the arrow is not important - it can be placed anywhere else by parallel transport.

The straight line, the plane, or the space around us are all instances of a *flat* space. The idea of a vector as an arrow extending from a chosen origin up to a given point works for such a flat space, but not in a *curved* one.

Consider a curved surface C shown in fig. 17-9 with a chosen 'origin' O, and a point P in it. The arrow extending from O to P does not make sense in the context of the surface itself (it does make sense in the context of the three dimensional space in which the surface is embedded, but only because that space is a flat one).

However, if we consider the *tangent* plane T at P, then that plane is a flat space and the

idea of a vector as an arrow works in it. Indeed, one can draw a lot of tangent vectors at P, all lying in the tangent plane and, in this way, can associate not one single vector but a whole lot of tangent vectors with the given point P, where all these tangent vectors make up a vector space. In the special case of the surface being a flat one, the radius vector associated with P is just one among all these tangent vectors, since the tangent plane T then coincides with the surface C itself.

In summary, given a point P in a curved space C, the idea of a directed line segment from a chosen origin O up to that point (the ‘radius vector’) is not a meaningful one while, instead, the idea of a *tangent space* at that point is. The tangent space can be defined in terms of *infinitesimal* tangent vectors drawn at P. In the case of a flat space C, the tangent space at O (or at any other point P) coincides with C itself, and the vector extending from O up to any other point P is just one among the vectors in T.

In the special relativistic theory, the space of events, represented by a (1+3)-dimensional space-time diagram with the help of space-time co-ordinates as measured in an inertial frame S is, as we have seen, a flat one. In this space, any point P with co-ordinates (ct, x, y, z) defines a vector, i.e., the radius vector extending from the origin $(0, 0, 0, 0)$ up to P. Choosing unit basis vectors along the space-time co-ordinate axes, the components of the radius vector are seen to be just the co-ordinates themselves, i.e., (ct, x, y, z) . These components get transformed under a Lorentz transform as in (17-4a), where the transformation matrix L is of the general form (17-17). These transformation formulae are characteristic of what are referred to as *four-vectors*.

In the *general* relativistic theory, however, the space-time events do not make up a flat space, and one needs the idea of the tangent space at any arbitrarily chosen point in order to define four-vectors. The tangents are defined in terms of the infinitesimal changes in the space-time co-ordinates $(c\delta t, \delta x, \delta y, \delta z)$ at the point under consideration, where these can have arbitrarily chosen (but infinitesimally small) values. With four-vectors defined by these infinitesimal changes, the transformation of the components of a four-vector is determined by the way these infinitesimal changes are transformed under a transformation of the co-ordinates.

What is important to mention here is that, the transformations of co-ordinates that one needs to consider in the general relativistic theory are of a much broader variety than the Lorentz transformations of the special relativistic theory. But these we will consider at a later point in this chapter. For now, we will look at the representation of physical quantities as four-vectors in the special relativistic theory.

17.2.8.2 Four-vectors

The special relativistic principle of equivalence tells us that physical laws have to have identical mathematical forms in all inertial frames of reference. The mathematical equation expressing a physical law generally relates a number of physical quantities with one another. For instance, Newton's second law relates the acceleration (i.e., the time derivative of the velocity) of a particle with the force acting on it. Now, the measure of velocity, acceleration or force depends on the frame of reference in which the motion of the particle is observed, and these get transformed when considered in some other frame of reference.

If the physical quantities appearing on the two sides of the mathematical equation expressing a physical law do not change in the same manner in the transformation from one inertial frame to another (i.e., in a Lorentz transformation), then that amounts to a violation of the special relativistic principle of equivalence.

A set of quantities that transform similarly under a Lorentz transformation is made up of the so-called *four-vectors*. This is analogous to the similarity in the transformation of the components of our familiar three dimensional vectors under a rotation of the Cartesian axes (refer to section 2.9), all of which transform in a manner similar to the components of the radius vector (x, y, z) of an arbitrarily chosen point. In the case of (1+3)-dimensional space-time, a four-vector is, similarly, an object specified by four components in any given inertial frame of reference, and the basic properties relating to vector addition and multiplication by scalars imply that these components follow the same transformation rules in a change of the frame of reference as the space-time co-ordinates (ct, x, y, z) of a point.

As I have already mentioned, this approach works in the special relativistic theory, but not in the general relativistic one, where a four-vector means an object made up of a set of four components (when considered with reference to any given co-ordinate system, of which more later) that transforms in a manner analogous to *infinitesimal changes* ($c\delta t, \delta x, \delta y, \delta z$) of the co-ordinates.

Incidentally, the *defining* property of a vector is not the way it transforms, but its membership in a linear vector space, which implies the transformation property as a consequence.

Finally, I have to mention here that, because of the curvature of space-time in the general relativistic theory (I again anticipate!), it is not sufficient to say that a physical quantity is a four-vector because one also needs to mention the point at which the four-vector is *based*, since the latter specifies the tangent space in which the four-vector resides.

With all this background out of our way, I can now give you the basic formulae relating to four-vectors and *tensors* as these appear in problems of physical relevance. For this, I will have to first explain to you a useful notation that turns out to be a greatly convenient one in writing out mathematical expressions in relativistic theory.

17.2.8.3 Four-vectors and tensors: a primer

First of all, we change the notation for space-time co-ordinates, as observed in an inertial frame, from ct, x, y, z to x^0, x^1, x^2, x^3 so as to underline the fact that these make up a single physical object, the *co-ordinate four-vector* (the radius vector in (1+3)-dimensional space). Together, these can be represented in the form of a column, as in (17-4a) or, in brief, as x^μ , with a Greek superscript whose value ranges from 0 to 3 (depending on the context, x^μ may stand either for a single component of the four-vector or may refer to the four-vector itself, with reference to some chosen frame S).

Other examples of four-vectors will be encountered in the following sections where, as mentioned above, the components of a four-vector are found to be transformed in a

change of the frame of reference in a manner similar to those of the co-ordinate four-vector

Consider an equation expressing a physical law in a frame of reference S, where it includes a four-vector A^μ as one of its terms. Then the other terms in the equation must also be four-vectors so that the equation remains unaltered in form when the same law is expressed in some other frame of reference S'.

Thus, an equation of the form

$$A^\mu = B^\mu \quad (\mu = 0, 1, 2, 3), \quad (17-51)$$

is in accord with the principle of equivalence since the four-vectors on the two sides transform similarly under a change of the frame of reference, both being similar to the transformation of the co-ordinate four-vector x^μ . Under a general Lorentz transformation as in fig. 17-2 (refer to sec. 17.2.4), x^μ gets transformed to x'^μ , where the latter stands for (ct', x', y', z') . Confining our attention, for the sake of simplicity, to transformation from a frame S to another frame S' related to S as in fig. 17-1, the components of a four-vector A^μ transform as

$$A'^0 = \gamma(A^0 - \beta A^1), \quad A'^1 = \gamma(A^1 - \beta A^0), \quad A'^2 = A^2, \quad A'^3 = A^3, \quad (17-52)$$

where, as before, the parameters β and γ are given by (17-2c), (17-4c), and where the superscripts are used to identify the components of the four-vector, and not to denote exponents. A^0 is referred to as the *time component* (or the *time part*) and the other three as the *space components* (the three together constituting the *space part*) of the four-vector A^μ .

In the relativistic theory, one works with mathematical objects that come with indices, where the indices may appear as superscripts (referred to as *contravariant* indices), or as subscripts (*covariant* indices), depending on what the objects stand for. Thus, x^μ and A^μ represent *contravariant four-vectors* in a given frame of reference S, while x'_μ and A'_μ

stand for the same contravariant four-vectors in another frame S' .

1. The term four-vector is commonly shortened to, simply, *vector*.
2. A statement like ' A^μ is a four-vector' may be a bit misleading since, more precisely, A^μ stands for the representative of the vector by its components in some particular frame of reference S , while A'^μ represents the same vector in some other frame S' . The vector itself is an abstract object (that can be denoted by the symbol A in the present instance) residing in a four dimensional vector space (the space of *events*; it is this space of events that is represented by means of a (1+3)-dimensional space-time diagram for any given frame of reference S). All this is left implied when one says that A^μ is a four-vector - more specifically, a contravariant one.

Associated with the contravariant vector A^μ , is the corresponding *covariant* vector A_μ , where the latter is obtained from the former by the process of *lowering* of the contravariant index, according to the formula

$$A_\mu = \sum_{\nu=0}^3 g_{\mu\nu} A^\nu \quad (\mu = 0, 1, 2, 3). \quad (17-53)$$

In this expression $g_{\mu\nu}$ is an object with two covariant indices, of which the index ν is summed over (which makes it a *dummy* index). The covariant index μ , on the other hand, remains *free*, and the summation on the right hand side produces the covariant vector A_μ .

The set of quantities $g_{\mu\nu}$ is one of basic relevance in the special relativistic theory, and is a *tensor* with two covariant indices (you need not worry what the term 'tensor' means; we will have a look at tensors presently; generally speaking, a tensor is specified with a number of indices, some of which may be contravariant and some others covariant; the *metric tensor* $g_{\mu\nu}$ has two indices, both of which we take, for now, to be covariant ones). It is this object that determines the geometry of the $(1 + 3)$ dimensional space of space-time events, and is defined with reference to the squared space-time separation (δs^2) between two neighbouring points, where the expression for the latter is of the general

form

$$\delta s^2 = \sum_{\mu\nu} g_{\mu\nu} \delta x^\mu \delta x^\nu. \quad (17-54)$$

1. The term '(1+3)-dimensional' is used to underline the fact that the time co-ordinate is not *quite* on the same footing as the space co-ordinates, since the principle of causality relates to the time co-ordinate and not to the space co-ordinates. However, this understood, we will refer to the space of events as a '*four dimensional*' one.
2. The above expression stands for the invariant squared separation between two neighboring points $\{x^\mu\}$ and $\{x^\mu + \delta x^\mu\}$ in a *semi-Riemann space*, of which the four dimensional space of events considered in the above paragraphs is a special instance. In the general case, the coefficients $g_{\mu\nu}$ depend on the co-ordinates $\{x^\mu\}$.

On comparing this general form with the expression (17-49) we find that the elements of the metric tensor make up the following 4×4 matrix,

$$(g_{\mu\nu}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \quad (17-55)$$

One observes that these do not depend on the co-ordinates $\{x^\mu\}$, which means that the four dimensional space of space-time events is a *flat* one. Moreover, the values of the elements $g_{\mu\nu}$ are the same in all reference frames (recall that the frame S is an arbitrarily chosen one). These elements make up a *diagonal* matrix, where the entries in the diagonal are $1, -1, -1, -1$, the same as the coefficients occurring on the right hand side of (17-49), which is only to be expected since the co-ordinates x^μ ($\mu = 0, 1, 2, 3$) are nothing but the old co-ordinates ct, x, y, z .

1. In the case of a flat space, where the coefficients $g_{\mu\nu}$ do not depend on the co-ordinates, the diagonal form holds at all points throughout the space. If, on the

other hand, these happen to depend on the co-ordinates (which is the same thing as saying that the space is curved), the matrix can, in general, be diagonalized at a single chosen point, but not at other points.

2. The coefficients $g_{\mu\nu}$ are the components of a tensor with two covariant indices. As with vectors, the tensors obey definite transformation rules, from which one can calculate the corresponding elements in some other frame of reference. As can be seen by referring to the transformation rules (see below), the transformed coefficients $g'_{\mu\nu}$ in the new frame form the *same* matrix that occurs on the right hand side of (17-55).

In addition to the metric tensor $g_{\mu\nu}$ with two covariant indices (we will call these the *lower* indices), one can have the tensor with two contravariant (*upper*) indices, and also the tensor with one upper and one lower index. Of these, the components of the tensor with two upper indices make up a 4×4 matrix *identical* with the right hand side of (17-55), i.e.,

$$\{g^{\mu\nu}\} = \{g_{\mu\nu}\}, \quad (17-56)$$

where the curly brackets denote the matrix formed of the relevant tensor components.

The element-wise identity of the metric tensor with two upper indices and the one with two lower indices follows from the general rule for raising and lowering of indices, of which (17-53) constitutes one instance, and (17-57) below is another.

The metric tensor $g^{\mu\nu}$ with two upper indices is useful for the purpose of *raising* of an index of a tensor in a manner analogous to the lowering of an index, as in (17-53). Thus, associated with a covariant vector A_μ , one has the corresponding contravariant vector A^μ given by

$$A^\mu = \sum_{\nu} g^{\mu\nu} A_\nu \quad (\mu = 0, 1, 2, 3), \quad (17-57)$$

where once again there is a summation over the dummy index ν . Making use of the above formula, one obtains the following simple relations between the components of

A^μ and A_μ

$$A_1 = A^1, A_2 = -A^2, A_3 = -A^3, A_4 = -A^4, \quad (17-58)$$

where similar relations hold for the components in any other frame of reference.

We are now in a position to introduce the concept of *tensors*. From the mathematical point of view, tensors are defined with reference to the linear vector space of vectors. Thus, if V denotes a linear vector space (for instance, the space of four-vectors with one single upper index) then a tensor (with two upper indices in this instance) is a mathematical object that resides in the *direct product* space $V \otimes V$, where the concept of the direct product is a basic one in set theory.

However, we need not concern ourselves with this abstract definition, since what works in practice is the characterization of tensors in terms of their transformation properties. For instance, a tensor $T^{\mu\nu}$ with two upper indices is a mathematical object that transforms in a manner similar to the product $A^\mu B^\nu$, where A^μ and B^ν denote any two four-vectors, each with one upper index. A tensor with two lower indices or a tensor with one upper and one lower index can be similarly characterized. It is in this sense that $g^{\mu\nu}$ and $g_{\mu\nu}$ were referred to as metric *tensors* with, respectively, two upper and two lower indices.

One can generalize to tensors with arbitrarily specified numbers of upper and lower indices, but this will not be needed for our purpose. The total number of indices in a tensor fixes its *rank*. Thus, a four-vector is a tensor of rank one, while the metric tensor is of rank two. Finally, a *scalar* is a tensor of rank zero, consistent with fact that it remains invariant in a space-time co-ordinate transformation.

In mathematical literature the term 'rank' of a tensor is used in a different sense. What has been termed 'rank' here is then referred to as the 'order' of a tensor.

Incidentally, the process of summing over a dummy index is an effective one not only in formulae like (17-53) and (17-57), where the metric tensor is used for the raising or lowering of indices, but also in formulae where the rank of a tensor gets changed. For

instance for given four-vectors A^μ , B_ν , the products $A^\mu B_\nu$ ($\mu, \nu = 0, 1, 2, 3$) constitute the components of a tensor of rank two, in which there are two free indices (i.e., ones that are not summed over), one contravariant and one covariant. If now, one sets $\mu = \nu$ and sums over the resulting single index (which now makes it a dummy index), one ends up with $\sum_\mu A^\mu B_\mu$ and, on invoking the rule of transformation of four-vectors, this can be seen to be a scalar. In other words, this process, referred to as a *contraction* reduces the rank of a tensor by two.

Indeed, the equations (17-53) and (17-57) can be looked upon as instances of contraction. For instance, consider the product $g_{\mu\nu} A^\sigma$ ($\mu, \nu, \sigma = 0, 1, 2, 3$), which stands for a tensor of rank three, with one upper index and two lower indices. Now set $\sigma = \nu$ and then sum over the dummy index ν . This is a contraction that reduces the rank by two and results in the covariant vector A_μ of (17-53).

Finally, for given four-vectors A^μ , B_ν , the contraction

$$\sum_\mu A^\mu B_\mu = A^0 B^0 - A^1 B^1 - A^2 B^2 - A^3 B^3, \quad (17-59)$$

which is a scalar, is referred to as the *scalar product* (or the *inner product*) of the two four-vectors. By contrast the product $A^\mu B_\nu$ is the *outer product* of the two vectors A , B and is a tensor of rank two.

From the mathematical point of view, the covariant vector A_μ can be looked upon as the *dual* of the contravariant vector A^μ , and the two vectors reside in vector spaces dual to each other. The inner product then signifies the scalar resulting from the action of the dual of A , on B .

In closing this section, I refer to the classification of four-vectors as *time-like* and *space-like* ones. A four-vector A^μ is termed a time-like one if the scalar $\sum_\mu A^\mu A_\mu$ has value greater than zero while, on the other hand, it is referred to as a space-like four-vector if the above scalar has a value less than zero. Finally, $\sum_\mu A^\mu A_\mu = 0$ corresponds to A^μ being a *light-like* four-vector.

Considering two space-time events with co-ordinates x^μ and $x^\mu + \delta x^\mu$ ($\mu = 0, 1, 2, 3$) (in some chosen inertial frame S), where the increments δx^μ are infinitesimal ones, being otherwise arbitrary, the space-time separation $\delta s^2 = \delta x^\mu \delta x_\mu$ can have a positive, negative, or zero value (i.e., the vector δx^μ can be time-like, space-like, or light-like). However, if we consider a moving particle with instantaneous velocity v as measured in S, and if the two events referred to above correspond to the particle being located at points (x^1, x^2, x^3) and $(x^1 + \delta x^1, x^2 + \delta x^2, x^3 + \delta x^3)$, at scaled times x^0 and $x^0 + \delta x^0$, then the space-time separation between the two events is necessarily a time-like one and accordingly, the vector δx^μ is also time-like.

17.2.8.4 The velocity four-vector

Looked at in any given inertial frame S, a four-vector is made up of a time component and three space components where, in a transformation to a new frame S' (which, for the sake of simplicity, we assume to be related to S as in fig. 17-1), the components transform as in (17-52).

In sec. 17.2.5.4, we derived the transformation properties of the three components of the velocity of a particle under a similar transformation of the frame of reference (eq. (17-32)). Making use of that transformation law, and defining the parameter γ as

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}, \quad (17-60)$$

analogous to the definition in eq. (17-2c), one arrives at the result that the four quantities

$$w^0 = \gamma c, \quad w^1 = \gamma v_x, \quad w^2 = \gamma v_y, \quad w^3 = \gamma v_z, \quad (17-61)$$

transform like the components of a four-vector as in (17-52) (check this out). Indeed, the quantities w^μ ($\mu = 0, 1, 2, 3$) are nothing but

$$w^\mu = \frac{\delta x^\mu}{\delta \tau}, \quad (17-62)$$

where δx^μ ($\mu = 0, 1, 2, 3$) are simply the infinitesimal changes $c\delta t, \delta x, \delta y, \delta z$ relating to the components of the displacement of the particle (in any chosen inertial frame S) in time δt , and $\delta\tau$ is the *proper time interval* corresponding to δt (check *this* out as well). Since $\delta\tau$ is an invariant quantity under a change of the frame of reference, i.e., is a scalar, and δx^μ is a four-vector (note that $\delta\tau$ is a meaningful quantity since δx^μ is time-like), w^μ is a four-vector as well - a time-like one. This is termed the ‘velocity four-vector’ (or, in brief, the *four-velocity*) of the particle, referred to the frame S

The three dimensional vector $\mathbf{w} = \gamma\mathbf{v}$, constituting the space part of w^μ is commonly referred to as the *proper velocity* of the moving particle, since it is the derivative of the displacement $\delta\mathbf{r}$ with respect to the proper time interval $\delta\tau$.

At times, it is more convenient to work with a dimensionless four-velocity u^μ defined as

$$u^\mu = \frac{1}{c}w^\mu. \quad (17-63)$$

It is straightforward to verify that u^μ (as also w^μ) is indeed a time-like four-vector, since

$$u^\mu u_\mu = 1, \quad (17-64)$$

(check this out).

In the above paragraphs I have, at times, referred to transformations between frames related to each other as in fig. 17-1. Similar results follow for Lorentz transformations of the general type considered in sec. 17.2.4.

Problem 17-7

Express the velocity \mathbf{v} in terms of the proper velocity \mathbf{w} (i.e., find the inverse of the relation

$$\mathbf{w} = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \mathbf{v}.$$

Answer to Problem 17-7

ANSWER: $\mathbf{v} = \frac{1}{\sqrt{1+\frac{w^2}{c^2}}} \mathbf{w}.$

Problem 17-8

For any given time-like four-vector A^μ , show that one can find a space-like four-vector B^μ such that $A^\mu B_\mu = 0$ (two four-vectors of which the inner product is zero, are said to be *orthogonal* to each other).

Answer to Problem 17-8

HINT: For a time-like four-vector, one can always find a Lorentz frame (i.e., an inertial frame of reference) in which its space components are all zero. Hence, consider the frame in which $A^i = 0$ ($i = 1, 2, 3$). The invariant inner product $A^\mu B_\mu$, when evaluated in this frame (call it S_0) evaluates to $A^0 B_0$. Now, in any given frame, one can always choose a space-like four-vector B^μ such that $B_0 (= B^0) = 0$, since any four-vector for which $B^{02} < \sum_i B^{i2}$ is, by definition, a space-like one (which means that any four-vector with at least one non-zero space component, can be chosen as B^μ , subject to $B^0 = 0$). With *this* choice for the components of B^μ in S_0 , one can transform to any other frame in which one will have $\sum_\mu A^\mu B_\mu = 0$, since this represents the invariant inner product.

17.2.8.5 Relativistic mass, relativistic momentum, and relativistic energy

Consider an inertial frame S in which a moving particle is observed to have an instantaneous velocity \mathbf{v} at time t , and also the frame S_0 in which the particle is instantaneously at rest (which makes S_0 the instantaneous *rest frame* of the particle).

The mass m_0 of the particle, as measured in S_0 , is termed its *rest mass*, and is one of its intrinsic properties. One important lesson of the special theory of relativity is that the ‘moving mass’ (see below) differs from the rest mass m_0 .

We consider, for the given frame S , the quantity

$$m = m_0 \gamma, \tag{17-65}$$

that pertains to the moving particle with instantaneous velocity \mathbf{v} , as observed in S, γ being defined as in (17-60). This is referred to as the *mass* of the particle with velocity v (at times, the term *moving mass* is also used so as to distinguish it from the rest mass, which is the limiting value of the moving mass for $v \rightarrow 0$; the term *relativistic mass* is also commonly employed).

From the physical point of view, m does indeed have the significance of the mass of the moving particle, as can be seen from various considerations, one of these considerations being that the vector $m\mathbf{v}(=m_0\mathbf{w})$ has the significance of the momentum of the particle (the *relativistic momentum*), looked at from the point of view of the observer in the frame S. Additionally, the expression $mc^2(=m_0\gamma c^2)$ can be interpreted as the relativistic *energy* of the moving particle. All this can be seen as follows.

Consider a closed system of two particles A, B, observed in S, that collide with each other and have velocities \mathbf{v}_A and \mathbf{v}_B just before the collision. The collision may result in the same two particles moving away with velocities that may differ from \mathbf{v}_A and \mathbf{v}_B , or in one or more particles differing from these two. We assume that there are two resulting particles that we call C and D, where C, D may be the same as A, B. Let the velocities of C, D, just after the collision, be \mathbf{v}_C , \mathbf{v}_D respectively.

Then, assuming that $m\mathbf{v}$ can indeed be interpreted as the momentum of a particle having velocity \mathbf{v} , *the principle of conservation of momentum*, as expressed by an observer in S, can be expressed in the form

$$m_A\mathbf{v}_A + m_B\mathbf{v}_B = m_C\mathbf{v}_C + m_D\mathbf{v}_D, \quad (17-66)$$

where m_A , m_B , m_C , m_D stand for the moving masses of A, B, C, D for their respective velocities as observed in S (thus, for instance, $m_A = \frac{1}{\sqrt{1-\frac{v_A^2}{c^2}}}m_{0A}$, where m_{0A} is the rest mass of A).

Additionally, with the above interpretation of the relativistic energy of a moving particle,

the principle of conservation of energy, as expressed in the frame S, is

$$m_A c^2 + m_B c^2 = m_C c^2 + m_D c^2, \quad (17-67)$$

Let us now consider the same collision from the point of view of some other frame S' where, for the sake of simplicity, one can assume that S' is related to S as in fig. 17-1. We denote the masses and velocities of the particles before and after the collision by attaching a prime to each of the respective symbols as referred to the frame S.

One then finds, by starting from (17-66), (17-67), and applying the velocity transformation formulae of sec. 17.2.5.4, that the momentum conservation equation holds in the frame S' as well:

$$m'_A \mathbf{v}'_A + m'_B \mathbf{v}'_B = m'_C \mathbf{v}'_C + m'_D \mathbf{v}'_D, \quad (17-68)$$

What is more, the principle of conservation of energy turns out to be valid in the frame S' as in S:

$$m'_A c^2 + m'_B c^2 = m'_C c^2 + m'_D c^2, \quad (17-69)$$

Recall that we have identified the expression $m\mathbf{v}$ with the momentum of a moving particle in the relativistic context, based on the interpretation of the expression $m = m_0\gamma$ as the relativistic mass, where m_0 stands for the rest mass. Additionally, we have identified the expression $mc^2 = m_0\gamma c^2$ as the relativistic energy of the particle. The consistency of these interpretations is then borne out by the result stated above: *the principles of conservation of energy and momentum are valid regardless of the frame of reference.*

The principles of conservation of energy and momentum are consequences of the homogeneity of time and space. The latter being general principles characterizing physical phenomena, the frame-independence of the conservation of relativistic energy and momentum conforms to the special relativistic equivalence principle.

This indicates that the expressions for the relativistic mass, momentum, and energy of a particle we started from are consistent with the principle of equivalence. All ex-

perimental evidence relating to motions of systems of particles corroborate that these expressions are indeed a set of consistent ones.

17.2.8.6 The energy-momentum four-vector

The rest mass m_0 is, by definition, a scalar quantity since its value does not depend on the frame of reference.

Given a particle moving with velocity \mathbf{v} as observed in a frame S, the S-bound observer assigns the value $\gamma^{-1}m$ to the rest mass, where $\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$. A similar expression holds for the rest mass in a second frame S'. But these quantities have the same value, since both are equal to the mass of the particle in its rest frame.

Hence, considering the inertial frame S in which the particle has velocity \mathbf{v} , the four quantities $p^\mu = m_0 w^\mu$ ($\mu = 0, 1, 2, 3$) constitute the components of a four-vector (a time-like one, since w^μ , the *velocity four-vector* (or the *four-velocity*), is time-like) in S, where the time component is $p^0 = mc = \frac{E}{c}$, and the three space components constitute the vector $\mathbf{p} = m\mathbf{v}$, E and \mathbf{p} being respectively the relativistic energy and the relativistic momentum of the particle in the frame S. This four-vector,

$$p^\mu = m_0 w^\mu, \quad (17-70a)$$

having components

$$p^0 = m_0 \gamma c, \quad \mathbf{p} = m_0 \gamma \mathbf{v} = m_0 \mathbf{w}, \quad (17-70b)$$

is referred to as the momentum four-vector (or the *four-momentum*; the term 'energy-momentum four-vector' is also used so as to indicate that it is a composite object incorporating both the energy and the momentum) of the particle.

Referring to the time component ($\mu = 0$) of the four-momentum in the above relations, it can be expressed, as mentioned above, as

$$p^0 = \frac{E}{c}, \quad (17-71a)$$

on using the formula for the relativistic energy (refer to sec. 17.2.8.5)

$$E = mc^2, \quad (17-71b)$$

that expresses the principle of *mass-energy equivalence*.

Noting the relation

$$m = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}}, \quad (17-72)$$

for the relativistic mass, one can expand the expression for the relativistic energy for a particle of given rest mass m_0 in powers of $\frac{v}{c}$ as

$$E = m_0c^2 + \frac{1}{2}m_0v^2 + \cdots, \quad (17-73)$$

where the first term is a constant contribution, referred to as the *rest energy*, and the second term is seen to be nothing but the non-relativistic expression for the kinetic energy of the particle. This is consistent with the fact that the higher order terms all tend to zero in the limit $\frac{v}{c} \rightarrow 0$ when compared with the non-relativistic kinetic energy term. Thus, the expression for the relativistic energy finds a consistent interpretation in the statement that the first term in the above expansion is in the nature of an internal energy, representing the energy of the particle in its rest frame, the second term represents the kinetic energy in the non-relativistic limit, while the remaining terms represent the relativistic corrections to the non-relativistic expression.

The *relativistic kinetic energy* of the particle is defined as

$$T = (m - m_0)c^2 = m_0(\gamma - 1)c^2, \quad (17-74)$$

which includes the non-relativistic contribution to the kinetic energy as also all the higher order relativistic corrections.

Finally, note that the invariant inner product $\sum_{\mu} p^{\mu} p_{\mu}$ formed with the four-momentum

p^μ , when evaluated in the frame S, has the value

$$\sum_{\mu} p^\mu p_\mu = \frac{E^2}{c^2} - p^2, \quad (17-75a)$$

while the same quantity, when evaluated in the rest frame S_0 of the particle, is

$$\sum_{\mu} p^\mu p_\mu = m_0^2 c^2, \quad (17-75b)$$

(check these statements out). Since these two expressions must be equal, one obtains the following relation between the relativistic energy and the relativistic momentum,

$$E^2 = p^2 c^2 + m_0^2 c^4. \quad (17-75c)$$

Problem 17-9

The non-relativistic kinetic energy of a particle of rest mass $m_0 = 1.2 \times 10^{-4} \text{ kg}$ is $T_0 = 2.5 \times 10^6 \text{ J}$. Find the relativistic correction to the kinetic energy and compare it with the non-relativistic kinetic energy.

Answer to Problem 17-9

SOLUTION: Since the non-relativistic expression for the kinetic energy is $T_0 = \frac{1}{2} m_0 v^2$, we have, $v^2 = \frac{2T_0}{m_0}$. At the same time, $\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \approx 1 + \frac{1}{2} \frac{v^2}{c^2} + \frac{3}{8} \frac{v^4}{c^4}$ (up to the second degree term in the small quantity $\frac{v^2}{c^2}$). Thus, the kinetic energy is (refer to formula (17-74)) $T \approx m_0 c^2 (\frac{1}{2} \frac{v^2}{c^2} + \frac{3}{8} \frac{v^4}{c^4})$, of which the first term ($\frac{1}{2} m_0 v^2$) is the non-relativistic part and the second term ($\delta T \approx \frac{3}{8} m_0 \frac{v^4}{c^2}$) gives the relativistic *correction* to the kinetic energy in the leading order of approximation. Thus, using the expression for v^2 in terms of T_0 we get, $\delta T \approx \frac{3}{2} \frac{T_0^2}{m_0 c^2}$, from which, using $c \approx 3 \times 10^8 \text{ m} \cdot \text{s}^{-1}$, and the given values of T_0, m_0 , we obtain, $\delta T \approx \frac{3 \times 2.5 \times 2.5 \times 10^{12}}{2 \times 1.2 \times 10^{-4} \times 3 \times 3 \times 10^{16}} \approx 0.87 \text{ J}$, and $\frac{\delta T}{T_0} \approx 3.48 \times 10^{-7}$.

17.2.8.7 The Döppler effect

Consider a plane monochromatic wave of light of angular frequency ω , with wave vector \mathbf{k} , as observed in an inertial frame of reference S. Then any of the Cartesian components

of the electric or the magnetic field vector (we call it η) can be expressed in the form

$$\eta = Ae^{i\psi(\mathbf{r},t)}, \quad (17-76a)$$

where the phase of the wave at the point \mathbf{r} at time t is of the form

$$\psi = \mathbf{k} \cdot \mathbf{r} - \omega t. \quad (17-76b)$$

Looked at from some other inertial frame S' , the field component undergoes a transformation (the field components make up a tensor of rank two, as we will find later), which means that the amplitude component A changes accordingly. However, the phase ψ is an *invariant*, which means that the wave vector and the frequency have to change as the wave is described in the frame S' , so as to make possible the equality

$$(\psi = \psi' :) \mathbf{k} \cdot \mathbf{r} - \omega t = \mathbf{k}' \cdot \mathbf{r}' - \omega' t'. \quad (17-77)$$

The invariance of the phase at any given space-time point is a consequence of the fact that the value of the phase can be made to possess a physical significance. For instance, one can form a *wave packet* by superposing waves having their frequencies and wave vectors spread over a narrow range so that the wave packet has an identifiable maximum, distinct from all other maxima, at, say $\psi = 0$. This particular value of the phase can then be made to represent a *signal*, which is a physical entity that must have the same relevance in all inertial frames of reference. The invariance of phase is consistent with all known facts about electromagnetic fields, an instance of which is the relation $c = \nu\lambda$ relating the frequency and wavelength of a monochromatic wave in free space.

The equality (17-77) can be satisfied if the four quantities

$$k^0 = \frac{\omega}{c}, k^1 = k_x, k^2 = k_y, k^3 = k_z, \quad (17-78)$$

form the components of a four-vector, so that the phase appears as the invariant inner

product

$$\psi = - \sum_{\mu} k^{\mu} x_{\mu}. \quad (17-79)$$

Let us, for the sake of simplicity, assume that the frame S' is related to S as in fig. 17-1. Then the transformation of k^{μ} has to be of the form (17-52). In particular, the transformation of the time-like component ($\mu = 0$) tells us that the angular frequency gets changed in a Lorentz transformation as

$$\omega' = \gamma(\omega - c\beta k_x), \quad (17-80)$$

If S represents the frame in which the source of light is at rest while S' is the frame of an observer recording the frequency of light, then the above formula relates the frequency recorded by the observer to the frequency of light emitted by the source. The phenomenon of the latter differing from the former is referred to as *Döppler effect*. In the special case when the wave emitted by the source propagates along the x -axis of either frame of reference (i.e., if $k_x = k = \frac{\omega}{c}$), the above relation assumes the simple form

$$\omega' = \sqrt{\frac{1-\beta}{1+\beta}} \omega, \quad (17-81)$$

where, according to this formula, ω' can be less than or greater than ω depending on whether the velocity of S is along the positive or the negative direction of the x -axis.

The formula (17-81) constitutes the mathematical expression of the so-called *longitudinal* Döppler effect. In contrast, the formula for ω' looks different when the velocity of the observer (frame S') relative to the source (frame S) is perpendicular to the direction of the wave vector of the plane wave emitted by the source. Thus, assuming that $k_x = k_z = 0$, $k_y = \frac{\omega}{c}$, one obtains

$$\omega' = \gamma\omega, \quad (17-82)$$

which is the formula for the *transverse* Döppler effect.

In the literature, the term ‘transverse Döppler effect’ is, at times, used in the case where the direction of motion of the observer relative to the source is perpendicular to

the direction of wave propagation in the frame of the *observer*.

Incidentally, the transformation of the four-vector k^μ gives not only the change in frequency, but the change in the direction of propagation as well, the latter being a consequence of the transformation of the spatial components of k^μ . This change in the direction of propagation due to a Lorentz transformation is precisely the phenomenon of relativistic aberration considered in sec. 17.2.5.5.

There is possible another approach, equivalent to the one considered above, for describing and explaining the D ppler effect, where one adopts the *quantum mechanical* point of view. In this approach, a plane monochromatic wave appears as a quantum mechanical *harmonic oscillator*, where a stationary state of the oscillator can be interpreted in terms of *photons*. Each photon is characterised by a relativistic energy $\hbar\omega$ and a relativistic momentum $\hbar\mathbf{k}$, where ω and \mathbf{k} are the angular frequency and the wave vector of the monochromatic wave under consideration.

The four-momentum of the photon is thus

$$p^\mu = \left\{ \frac{\hbar\omega}{c}, \hbar\mathbf{k} \right\}, \quad (17-83)$$

and the D ppler effect is then seen as the transformation of the four-momentum resulting from a Lorentz transformation.

We had a look at the D ppler effect for acoustic waves in chapter 9. The D ppler effect for light differs in a number of ways from that of acoustic waves. First of all, for given values of ω and \mathbf{k} , the change in the frequency in the case of light depends only on the velocity of the observer relative to the source while, in the case of acoustic waves it depends additionally on the velocity of the medium in which the wave is set up. Secondly, the D ppler effect for acoustic waves is of the longitudinal type, while that in the case of light (or electromagnetic waves) can be of the longitudinal or transverse type or, generally speaking, can be a combination of the two. Finally, even when one considers the longitudinal D ppler effect in a stationary medium, the change in frequency, as given by

formula (17-81), differs from the corresponding formula for acoustic waves. All this goes to show that electromagnetic waves are fundamentally different in nature as compared to acoustic ones.

17.2.8.8 The force four-vector

The equations of motion of a particle in Newtonian mechanics, in any given inertial frame S, is

$$\frac{d\mathbf{p}^{[N]}}{dt} = \mathbf{F}^{[N]}, \quad (17-84)$$

where $\mathbf{p}^{[N]} (= m_0 \mathbf{v})$, $\mathbf{F}^{[N]}$ stand for the Newtonian momentum and the Newtonian force respectively, referred to the frame S. This equation is not form-invariant in a relativistic transformation from one inertial frame to another, and so needs modification in the relativistic context. One can express the dynamics by means of a form-invariant equation of motion involving four-vectors p^μ and F^μ , where p^μ is the four-momentum introduced in sec. 17.2.8.6 and F^μ is referred to as the *four-force*. This equation of motion, valid in the relativistic context, reads

$$\frac{dp^\mu}{d\tau} = F^\mu, \quad (17-85)$$

in which both sides are four-vectors (recall that $\delta\tau$, an infinitesimal increment in the proper time, corresponding to an increment δt of time measured in the chosen inertial frame S, is a scalar). Making use of the definition of the four-momentum p^μ in terms of the Newtonian momentum $\mathbf{p}^{[N]}$ ($p^\mu = (\frac{E}{c}, \gamma \mathbf{p}^{[N]})$; the relativistic energy E , determining the time-like component of p^μ can also be expressed in terms of the Newtonian momentum), one can work out the components of the four-force in terms of the Newtonian force $\mathbf{F}^{[N]}$, as I outline below.

In contrast to the Newtonian momentum $\mathbf{p}^{[N]} = m_0 \mathbf{v}$, the space part of p^μ (i.e., $m_0 \gamma \mathbf{v}$) is the relativistic momentum which we have denoted by the symbol \mathbf{p} . For the sake of clarity, and in conformity with the notation adopted below, one may also use the symbol $\mathbf{p}^{[R]}$. We will, however, stick with the simpler symbol \mathbf{p} .

Considering, first, the space components of (17-85), the three space components of the

four-force are obtained as

$$F^i = \frac{dp^i}{d\tau} = \gamma \frac{dp^i}{dt} = \gamma \frac{d}{dt}(\gamma m_0 v^i) \quad (i = 1, 2, 3), \quad (17-86)$$

where the index i indicates a spatial component, and where m_0 , the rest mass, is identified as the Newtonian mass of the particle under consideration.

The expression

$$\mathbf{F}^{[R]} = \frac{d\mathbf{p}}{dt}, \quad (17-87)$$

is referred to as the *relativistic* force; on multiplying this with γ , one obtains the space part of the four-force F^μ .

Noting the dependence of γ on the velocity v of the particle ($\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$), and working out the time derivative, one obtains the relativistic force (a three-vector) as

$$\mathbf{F}^{[R]} = \gamma m_0 (\mathbf{a}^{[N]} + \frac{\gamma^2}{c^2} (\mathbf{v} \cdot \mathbf{a}^{[N]}) \mathbf{v}) = \gamma (\mathbf{F}^{[N]} + \frac{\gamma^2}{c^2} (\mathbf{v} \cdot \mathbf{F}^{[N]}) \mathbf{v}), \quad (17-88a)$$

where

$$\mathbf{a}^{[N]} = \frac{d\mathbf{v}}{dt}, \quad (17-88b)$$

denotes the Newtonian acceleration (check eq. (17-88a) out). Thus, the relativistic force differs from the Newtonian force, but reduces to the latter in the non-relativistic limit.

The time component of the four-force

$$F^0 = \frac{dp^0}{d\tau} = \gamma \frac{dp^0}{dt} = \frac{\gamma}{c} \frac{dE}{dt}, \quad (17-89)$$

can also be similarly evaluated, and gives

$$F^0 = \frac{\gamma}{c} \mathbf{F}^{[R]} \cdot \mathbf{v} = \frac{\gamma^4}{c} \mathbf{F}^{[N]} \cdot \mathbf{v}, \quad (17-90)$$

(check this out).

Thus, finally, the relativistic equation of motion (17-85), which is form-invariant under a Lorentz transformation (since it contains a four-vector on either side) gives us the rates of change of the relativistic momentum and the relativistic energy, in terms of the Newtonian force as

$$\left(\frac{\gamma}{c} \frac{dE}{dt}, \gamma \frac{d\mathbf{p}}{dt}\right) = F^\mu, \quad (17-91a)$$

where

$$F^\mu = \{F^0, \mathbf{F}\} = \left\{\frac{\gamma}{c} \frac{dE}{dt}, \gamma \mathbf{F}^{[R]}\right\} = \left\{\frac{\gamma^4}{c} \mathbf{F}^{[N]} \cdot \mathbf{v}, \gamma^2 (\mathbf{F}^{[N]} + \frac{\gamma^2}{c^2} (\mathbf{v} \cdot \mathbf{F}^{[N]}) \mathbf{v})\right\}. \quad (17-91b)$$

One can express these four-force components in terms of the relativistic momentum \mathbf{p} instead of the velocity \mathbf{v} as

$$F^\mu = \{F^0, \mathbf{F}\} = \left\{\frac{\gamma^3}{m_0 c} \mathbf{F}^{[N]} \cdot \mathbf{p}, \gamma^2 (\mathbf{F}^{[N]} + \frac{1}{m_0^2 c^2} (\mathbf{p} \cdot \mathbf{F}^{[N]}) \mathbf{p})\right\}, \quad (17-92a)$$

where now γ is to be expressed in terms of p as

$$\gamma = \sqrt{1 + \frac{p^2}{m_0^2 c^2}}. \quad (17-92b)$$

If one knows the expression for the Newtonian force in some chosen frame of reference, then one can obtain the four-force in any arbitrary inertial frame by means of (17-91b) or (17-92a).

The relativistic equation of motion is made up of four component equations, of which the time component ($\mu = 0$) gives the rate of change of the relativistic energy in terms the work performed on the particle by the force acting on it (as we have seen above, this can be expressed in terms of the Newtonian force and the Newtonian velocity), while the space components give the rate of change of the relativistic momentum in terms of the relativistic force (which, once again, can be expressed in terms of the Newtonian force as in (17-88a)).

Notation and nomenclature in relativistic mechanics can, at times be a source of con-

fusion. For instance, one has to distinguish between the Newtonian force ($\mathbf{F}^{[N]}$ in our notation), the relativistic force ($\mathbf{F}^{[R]}$, eq. (17-88a)), and the spatial part of F^μ , the four-force, made up of components F^i ($i = 1, 2, 3$) where the three together have been denoted by \mathbf{F} above. In the case of velocity, the Newtonian velocity and the relativistic velocity are the same ($\frac{d\mathbf{r}}{dt}$), while the spatial part of the four velocity w^μ is obtained from either of the two by multiplying with γ . In the case of momentum, on the other hand, the relativistic momentum (\mathbf{p} in our notation) differs from the Newtonian momentum ($\mathbf{p}^{[N]}$) by a factor of γ , while the spatial part of the four-momentum is the same as the relativistic momentum. As for the mass, the relativistic mass (m) differs from the Newtonian mass (i.e., the rest mass) m_0 by the factor γ , while, similarly, the relativistic energy E differs from the rest energy by the factor γ . The *non-relativistic energy*, however, is not the same thing as the rest energy. The rate of change of the non-relativistic energy is given by $\mathbf{F}^{[N]} \cdot \mathbf{v}$, while the corresponding rate for the relativistic energy is $\mathbf{F}^{[R]} \cdot \mathbf{v}$ (check against (17-91b)).

17.2.9 The electromagnetic field as a tensor

The special theory of relativity is based on the special relativistic principle of equivalence, which tells us that the mathematical expressions of physical principles or laws must look similar in various different inertial frames which, in turn, implies that these mathematical formulae are to be expressed as equalities of *tensors* of identical ranks, since tensors of the same rank transform identically in a Lorentz transformation.

The relativistic equation of motion of a particle under the action of a force (eq. (17-85)) is a mathematical equation of this type, since it expresses the equality of two four-vectors, and constitutes the correct relativistic generalization of the Newtonian equation of motion (eq. (17-84)), the two sides of which do not have well defined and identical Lorentz transformation properties.

Considering a particle moving in an electromagnetic field, it turns out that its equation of motion can be expressed in the form (17-85) provided that the components of the electric and magnetic field vectors are assumed to make up a *tensor of rank two*, termed

the *electromagnetic field tensor* ($F^{\mu\nu}$) as

$$\begin{aligned} F^{0i} &= -\frac{E_i}{c} \quad (i = 1, 2, 3), \quad F^{12} = -B_3, \quad F^{13} = B_2, \quad F^{23} = -B_1, \\ F^{\mu\nu} &= -F^{\nu\mu} \quad (\mu \neq \nu) \quad (\mu, \nu = 0, 1, 2, 3), \end{aligned} \quad (17-93a)$$

which makes $F^{\mu\nu}$ an *antisymmetric* tensor. In these formulae, E_i, B_i ($i = 1, 2, 3$) stand for the components of the electric and magnetic field vectors. The tensor components $F^{\mu\nu}$ constitute the elements of a 4×4 antisymmetric matrix as

$$\{F^{\mu\nu}\} = \begin{pmatrix} 0 & -\frac{E_1}{c} & -\frac{E_2}{c} & -\frac{E_3}{c} \\ \frac{E_1}{c} & 0 & -B_3 & B_2 \\ \frac{E_2}{c} & B_3 & 0 & -B_1 \\ \frac{E_3}{c} & -B_2 & B_1 & 0 \end{pmatrix}. \quad (17-93b)$$

An implied assumption here is that *the charge of a particle is an invariant* under Lorentz transformations, which is consistent with experimental observations as also with all theoretical formulae.

It transpires from the above considerations that the electric and magnetic field strengths are not mutually independent entities, i.e., in other words, they make up a single composite entity, namely, the *electromagnetic field* or, more precisely, the *electromagnetic field tensor*.

Equipped with the electromagnetic field tensor, one can express the equation of motion of a particle of charge q in an electromagnetic field in the expressly covariant form (i.e., a form where all terms transform identically in a Lorentz transformation)

$$\frac{dp^\mu}{d\tau} = qF^{\mu\nu}w_\nu, \quad (17-94)$$

where the right hand side is identified as the electromagnetic four-force acting on the particle.

This is consistent with the familiar equation of motion involving the *Lorentz force*

$$\frac{d\mathbf{P}}{dt} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}), \quad (17-95)$$

in which the left hand side is the rate of change of the *relativistic* momentum.

Strictly speaking, this equation does not possess a meaningful *Newtonian* limit since, in this limit, the magnetic force $\mathbf{v} \times \mathbf{B}$ drops out.

While the above equation of motion tells us how the electromagnetic field affects the motion of a charged particle, one also needs to know *how the particle affects the field*. The name 'field' means that there is a tensor associated with every point in space-time, since one can associate an electric and a magnetic field with every point and every instant of time. *Electromagnetic theory* is built upon the basic idea that the field itself, made up of the electric and magnetic field strengths at all points in space, is a dynamical system, and the motion of a particle or a system of particles in an electromagnetic field is a consequence of an interaction of this dynamical system with the particle(s) under consideration.

Thus, the dynamics of the total system made up of the field on the one hand and of the particle(s) on the other is to be described in terms of the equation of motion of the particle(s), *and* that of the field. Even in the absence of charged particles, the dynamics of the field shows up in the form of space-time variations of the *free* electromagnetic field. These, precisely, are the *Maxwell equations* in free space in the absence of charges and currents, where these equations can now be written in a covariant (i.e., form-invariant) form by invoking the field tensor (along with the *dual* of the field tensor - a tensor field analogous to $F^{\mu\nu}$ and defined in terms of the latter, the two fields having distinct transformation properties under spatial inversion).

In the presence of charges and currents, one can define a *charge density* and a *current density* at every point in space and at every time instant. Together, these make up a four-vector field referred to as the *charge-current density*. One can then write down

the equation of motion of the field in a covariant form, involving the field tensor and the charge-current density four-vector, which turns out to be precisely the Maxwell equations in the presence of charges and currents.

From a fundamental point of view, the forces acting on a particle can be of electromagnetic origin or else, these may be in the nature of strong, weak, and gravitational forces. Of these, the gravitational ‘force’ is of a distinct nature since it appears in the theory not as a ‘true’ force but as a modification of the structure of space-time whereby the latter becomes a *curved* one. For weak fields, the motion of a particle in such a curved space-time appears to be analogous to the motion in a flat space-time under a force which is commonly referred to as the gravitational force.

This leaves us with the strong and weak forces that can possibly affect the motions of particles. However, these forces can be incorporated in a consistent theory only when the fields corresponding to these forces are considered from a *quantum* point of view. Strictly speaking, the electromagnetic field is also to be looked at as a quantum mechanical system, since the classical theory, outlined above, describing the interaction of charged particles with the electromagnetic field, leads to a number of inconsistencies, even though the theory can be formulated in terms of covariant equations.

An attempt at a consistent description of the strong, weak, and electromagnetic interactions of fundamental particles thus takes us to the domain of *quantum field theory*, which we will not enter into in this book. However, even accepting that a fundamental theory of interactions has to be in the nature of a quantum field theory, one can describe a great number of observed phenomena involving the electromagnetic interaction of charged particles by means of the relativistic classical theory of charged particles and electromagnetic fields outlined above in this section.

I will now present to you a number of basic ideas relating to the motion of particles in a gravitational field in the relativistic context, where the field is not necessarily weak and *interacts* with the particles. This constitutes the subject matter of the *general* theory of relativity. In this, I will again confine myself to the *classical* point of view since a

consistent *quantum* theory incorporating gravitation is yet to emerge.

17.3 The general theory of relativity: a brief introduction

17.3.1 Introduction: the general principle of equivalence

The special theory of relativity is based on a broadening of the Newtonian view of space and time describing the motions of systems of particles, so as to describe in a consistent manner the dynamics of particles *and* electromagnetic fields, where gravitational fields are either absent or sufficiently weak.

The general theory of relativity involves, in turn, a further broadening of the special relativistic view where a gravitational field is seen as a modification of the global metrical structure of the four dimensional space of all possible events while locally the structure remains that of the $(1 + 3)$ -dimensional Minkowski space.

The basic idea in the general theory comes from the observation that the motion of a free particle in a gravitational field is similar to the motion as observed in an *accelerated* frame in the absence of the field since, in either of these motions, *the acceleration is independent of the mass of the particle*. Here the term ‘free particle in a gravitational field’ means a particle on which the only influence is that of gravitation, with no other force acting on it.

As indicated in sections 3.10.3, 3.21, the acceleration of a particle in a frame of reference accelerated with respect to an inertial frame is independent of its mass, the inertial force arising due to the acceleration of the frame being proportional to the mass. While, in these sections, we considered only a uniform acceleration of the frame in translational motion and a pure rotational motion of the frame respectively, the conclusion remains true in general.

Considering a frame attached to a particle in free fall in a gravitational field, any other

particle (we refer to the latter as a 'test particle'), also in free fall but sufficiently close to the first particle, undergoes a uniform motion since the gravitational field strengths at the locations of the two particles being equal (recall that the two locations have been assumed to be close to each other), they have equal accelerations.

The term 'free fall' once again means motion in a gravitational field without any other influence acting on the particle under consideration.

In other words, a frame of reference attached to a particle in free fall in a gravitational field can be looked upon as an inertial frame as long as events close to the space-time location of that particle are considered

It is not sufficient to consider only those events whose spatial locations are close to the location of the reference particle under consideration since the time interval is also of importance here. For sufficiently large time intervals, the strength of the gravitational field may get altered and the test particle may acquire an acceleration with respect to the reference particle.

Hence, in describing the events in a small space-time region (R ; refer to fig. 17-10) around the space-time location of the reference particle one may make use of co-ordinates defined by an inertial frame. In this small space-time region, the special relativistic principle of equivalence holds, and the mathematical relations describing physical principles appear unchanged when new co-ordinates generated by a Lorentz transformation are used. However, now the admissible transformations are to be *specific* for the space-time location under consideration.

Considering a small region of space-time (R' ; see fig. 17-10) further away, a freely falling test particle (i.e., a test particle in free fall under the gravitational field in that region) will, in general, undergo an accelerated motion relative to the reference particle in R , and hence the inertial frame co-ordinates used for R will not be relevant for R' . In particular, the mathematical formulae expressing physical principles governing events

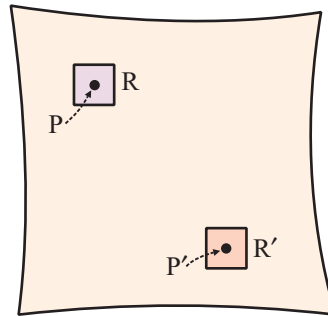


Figure 17-10: Portraying the universe of space-time events as a mosaic of small patches (schematic), such as the regions R , R' around the points P , P' ; local Lorentzian co-ordinates can be made use of in describing events in each such patch, where a patch can be identified as the tangent space at a point within it, the tangent space being made up of vectors based at that point.

in R' will not be form-invariant under the Lorentz transformations admissible for R . On the other hand R' will have its own group of Lorentz transformations under which these equations will remain form-invariant.

In other words, the universe of all possible space-time events can be looked upon as a mosaic of small patches like R , R' in fig. 17-10, where each of these small patches can be described by means of local co-ordinates in terms of which the metric $g_{\mu\nu}$ appears as in (17-55). There is no single co-ordinate system in terms of which the entire space of all possible events is characterized by a metric of this form, i.e., the entire space cannot be described by means of co-ordinates defined by one single inertial frame.

An inertial frame is, after all, nothing more than a convenient means of assigning co-ordinates to events, these co-ordinates being in the nature of a consistent set of book-keeping entries for the events. Given a co-ordinate system defined in terms of an inertial frame, one can make a transformation of some general kind so that the transformed co-ordinates are no longer tied to an inertial frame. Thus, even as the idea of a global inertial frame loses validity, that of a global co-ordinate system remains meaningful where such a co-ordinate system consistently attaches a set of four number (x^0, x^1, x^2, x^3) to each and every space-time event.

Thus, in the general theory of relativity, one can use an arbitrarily chosen co-ordinate

system, labeling each event with co-ordinates (x^0, x^1, x^2, x^3) such that, referring to events in any small patch like R, there exists a co-ordinate transformation by means of which one arrives at a local inertial frame co-ordinates wherein the metric assumes the form as in (17-55). There exists a set of local Lorentz transformations for which the special relativistic principle of equivalence holds, i.e., the equations expressing physical principles pertaining to events in R appear in a covariant form under these transformations.

17.3.2 Tensor fields

As mentioned above, a gravitational field leaves intact the local structure of space-time but distorts the global structure, which is in the nature of a mosaic of small patches, each having the structure of a (1+3)-dimensional Minkowski space. Globally, the space of events can be described in terms of an arbitrarily chosen co-ordinate system with co-ordinates (x^0, x^1, x^2, x^3) , where one can equally well choose any other co-ordinate system with co-ordinates (x'^0, x'^1, x'^2, x'^3) obtained by a co-ordinate transformation of an arbitrary choice.

The term 'arbitrary' in respect of co-ordinate systems and co-ordinate transformations, is to be qualified by a set of general requirements that are commonly left implied. Thus, a co-ordinate system has to satisfy certain continuity requirements as also the requirement that each possible event has to have a unique set of co-ordinates identifying it. Similarly, a co-ordinate transformation is to be a continuous and invertible one, in addition to possessing appropriate differentiability properties.

Referring to a co-ordinate system with co-ordinates $\{x^\mu\}$, the geometry of the space of events is described by a metric tensor $g_{\mu\nu}(x)$ that now varies from point to point in this space. For any given point P, one can find a co-ordinate transformation such that, in the new co-ordinates, the metric tensor assumes the form as in (17-55) in a small patch around that point. In mathematical terms, such a 'patch' is made up of infinitesimal increments $(\delta x^0, \delta x^1, \delta x^2, \delta x^3)$ or, in other words, is the *tangent space* at the point under consideration. One can choose co-ordinates that make explicit the fact that this tangent space has the structure of the (1+3)-dimensional Minkowski space of special relativity.

When one says that the various physical quantities, including the metric tensor, constitute tensor fields, one means that these correspond to tensors *based* at each specified point, where a tensor based at the point $x \equiv \{x^\mu\}$ gets transformed in a well defined manner under a *local* transformation, i.e., one in which the infinitesimal increments $\{\delta x^\mu\}$ get transformed to $\{\delta x'^\mu\}$. This transformation can be expressed in the form

$$\delta x'^\mu = \sum_{\nu} \frac{\partial x'^\mu}{\partial x^\nu} \delta x^\nu, \quad (17-96)$$

in which the *partial derivatives* of the transformed co-ordinates x' in terms of the original co-ordinates x are involved, where these partial derivatives now play the role of the elements of the matrix L of a Lorentz transformation, as in (17-17). The four infinitesimal quantities $\{\delta x^\mu\}$ now constitute a contravariant 4-vector based at x , while $\delta x'^\mu$ stands for the transformed four-vector. Any contravariant vector A^μ based at x then gets transformed to A'^μ as in (17-96). If there be such a vector based at every point x (subject to continuity and differentiability conditions), then that makes it a vector *field* $A^\mu(x)$, and the transformation formula at the point x reads

$$A'^\mu(x') = \sum_{\nu} \frac{\partial x'^\mu}{\partial x^\nu} A^\nu(x). \quad (17-97)$$

Tensor fields with arbitrary numbers of upper and lower indices can be defined from the the basic transformation formula for vectors, where the key role is played by the *metric tensor* field. Thus, the metric tensor $g_{\mu\nu}(x)$ with two lower indices produces a covariant tensor field from a contravariant one as

$$A_\mu(x) = \sum_{\nu} g_{\mu\nu}(x) A^\nu(x), \quad (17-98)$$

where $g_{\mu\nu}(x)$ gets transformed under a co-ordinate transformation like the product of two covariant vector fields. Generally speaking, a tensor field with m number of upper indices and n number of lower ones is referred to as one with rank (m, n) .

17.3.3 Einstein's equation for the metric tensor

The metric tensor is the object of central interest in problems in the the general theory of relativity. A gravitational field is completely characterized by the way it affects the geometry of space-time (i.e., by the way it causes the structure of space-time to deviate from that of a flat Minkowski space) and this, in turn is determined by the distribution of energy-momentum (of matter and electromagnetic radiation) and of the way the *flow* of energy-momentum is distributed in space and time. The momentum distribution, which relates to the flow of mass distribution, and the flow of energy-momentum are involved here because these determine the way the gravitational field varies in space *and* time.

The central equation determining the metric tensor in the general theory of relativity is *Einstein's equation*, which is a partial differential equation for $g_{\mu\nu}$ with a *source term* determined by the energy-momentum distribution and the distribution of the flux of energy-momentum. All these distributions are collected together in a single tensor field of rank two, referred to as the *stress-energy tensor*. It is the stress-energy tensor that enters into Einstein's equation as the source term, much like the charge-current density enters into Maxwell's equations as the source term for the electromagnetic field tensor (refer to sec. 17.2.9).

The energy term in the energy-momentum distribution can be taken to be the mass distribution by virtue of the mass-energy equivalence expressed in (17-71b).

Once the stress-energy tensor is known throughout space-time one can, in principle, determine the metric tensor by solving the Einstein equation. This, however, is not an easy job since the actual equation is of a complex structure, and relates a *non-linear* function of the metric tensor, known as the *Einstein tensor* $G_{\mu\nu}$, to the stress-energy tensor $T_{\mu\nu}$ as

$$G_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}, \quad (17-99)$$

where G stands for the Gravitational constant. The Einstein tensor $G_{\mu\nu}$ is related to the metric tensor $g_{\mu\nu}$ by means of a chain of intermediate formulae, starting from the

expression for the *Riemann curvature tensor*, which is a tensor of rank four defined in terms of the so-called *Christoffel symbols*. The Christoffel symbols depend non-linearly on the metric tensor and involve partial derivatives of the latter, and are useful objects for defining the *covariant derivatives* of tensor fields.

If one performs a straightforward partial differentiation on a tensor field, one does not end up with a tensor field of a higher rank, since the partial derivatives do not have the transformation properties of tensors. On the other hand, adding an extra term to the partial derivative involving the Christoffel symbol does produce a tensor field, which is how a *covariant derivative* is defined. All this has to do with the non-zero *curvature* of the space one is working with.

The Riemann curvature tensor tells us a great deal about the geometry of curved space-time, including all information relating to the curvature (or, more precisely, the curvatures) at any chosen point. Finally, two other objects necessary to define the Einstein tensor are the *Ricci tensor* (a tensor of rank two) and the *Ricci scalar*. These two, defined from the Riemann tensor by means of contraction of indices, contain progressively diminishing information relating to the geometry of space-time, but are still important in indicating the degree of curvature of space-time.

The Einstein tensor, resulting from the Ricci tensor and the Ricci scalar is the object of final interest that tells us how the energy-momentum distribution and energy-momentum flux in the space-time universe determines the curvature of space-time, where the curvature is the geometrical feature that indicates the ‘strength’ of the gravitational field at any given space-time point.

More specifically, it is the dependence of the Einstein tensor on the metric tensor $g_{\mu\nu}(x)$ that leads to the generalization of the concept of the ‘strength’ of the gravitational field at the point under consideration. The latter emerges naturally in the *non-relativistic* limit of the Einstein equation, which is the limit corresponding to a gravitational field of vanishingly small strength and to a vanishingly small rate of variation of the field. In this limit, one can choose co-ordinates such that the metric tensor deviates to only a

small extent from the Minkowski metric. Choosing an appropriate form of this metric, characterized by a small parameter Φ that measures the deviation from the Minkowski metric, one finds that the $\mu = 0, \nu = 0$ component of Einstein's equation (17-99) reduces to the form

$$\nabla^2 \Phi = 4\pi G \rho, \quad (17-100)$$

which is nothing but the *Poisson equation* in the Newtonian theory of gravitation, where the gravitational potential Φ is related to the mass density ρ , i.e., to the distribution of matter in space. In other words, the deviation of the metric tensor from the Minkowski metric given by (17-55) reduces, in the non-relativistic limit, to the gravitational potential, whose gradient gives the gravitational field strength. All this can be explicitly worked out in the simple case of a spherically symmetric gravitating body (refer to sec. 17.3.6).

17.3.4 Equation of motion in a gravitational field

The gravitational field (or, more precisely, the metric tensor) is determined by the mass and momentum distribution of matter and energy in space-time, and their fluxes. In the non-relativistic limit, this corresponds to the determination of the gravitational potential by the mass distribution. This, however, is only one side of the story, the other side being the way the flow of mass distribution (or, more generally, the stress-energy tensor) is determined by the gravitational field extending in space and time. In the non-relativistic limit, and for a weak gravitational field, this corresponds to the equation of motion of a particle or a system of particles in a gravitational field.

In other words, one has to consider the *composite system* made up of particles (or, equivalently of the mass distribution) and the gravitational field, where each of the two constituents of the composite system determine the space-time distribution of the other.

The relativistic generalization of the Newtonian equation of motion of a particle in a gravitational field is of the form of eq. (17-85)

$$m_0 \frac{d^2 w^\mu}{d\tau^2} = F^\mu, \quad (17-101)$$

where the four-force F^μ can be expressed in terms of the four-velocity w^μ and the Christoffel symbols, the latter depending on the metric tensor $g^{\mu\nu}$. The parameter τ can be related to an appropriately defined ‘proper time’ measured in the frame of the particle under consideration.

The basic idea is that the gravitational field expressed in terms of the metric tensor, or the curvature of space-time, is a dynamical system in its own right just as the electromagnetic field is. The analogy between the gravitational field and the electromagnetic field is a revealing one, where the Einstein equations (I use the plural to indicate that there are more than one components of the tensor form of the equation) correspond to the Maxwell equations in the presence of a charge-current distribution while the equation (17-101), with the appropriate expression substituted for F^μ , corresponds to the equation of motion of a charged particle (the so-called Lorentz equation) in an electromagnetic field.

This basic idea implies the existence of *gravitational waves*, analogous to electromagnetic waves implied by the Maxwell equations, where a gravitational wave consists of oscillations in the space-time curvature that propagate through space. It is not an easy matter to experimentally detect these waves, though recent observations appear to confirm their existence.

In concluding this section, I mention that, analogous to the equation of motion of a particle in a gravitational field, one can set up the equation of motion of a *photon* as well, where the four-momentum $m_0 w^\mu$ of the particle is to be replaced with the four-momentum $(\hbar\omega, \hbar\mathbf{k})$ of the photon. This approach tells us how the trajectory of a photon is affected by a gravitational field. Indeed, the curvature of light rays in strong gravitational fields has been experimentally observed.

17.3.5 Gravitation and the electromagnetic field

Gravitation couples not only to matter through the density and flux of mass-energy and momentum (the stress-energy tensor) of a system of particles (such a system, in the limit of continuously distributed matter, constitutes a ‘fluid’), but to electromagnetic

fields as well since, in principle, an electromagnetic field also contributes to the stress-energy tensor, thereby getting involved in the Einstein equations. On the other hand, gravitation affects the dynamics of an electromagnetic field since the Maxwell equations get modified in curved space-time. In the special relativistic flat space-time, the Maxwell equations can be cast straightaway into a covariant form by the simple expedient of using the electromagnetic field tensor, in which case the covariant form of the Maxwell equations involves the partial derivatives of the field tensor.

In a curved space-time, however, the partial derivatives do not form tensors whereas the *covariant derivatives*, defined with the help of the Christoffel symbols, do constitute tensors. Maxwell equations in the presence of gravity therefore are to be written with these covariant derivatives replacing the partial derivatives. In the process, the Maxwell equations, originally written in the context of flat space-time, get modified where the modified form now involves the metric tensor that enters into the definition of the Christoffel symbols.

17.3.6 The Schwarzschild solution

A simple but vastly important and interesting exact solution to Einstein's equations in general relativity was found by Schwarzschild, who worked out the metric tensor for a *spherically symmetric* gravitating body, where the gravitational field (or more precisely, the metric tensor) outside the body is a time independent one.

We Consider a spherically symmetric, static gravitating body of finite radius R , and start from the Minkowski metric of flat space-time

$$\delta s^2 = c^2 \delta t^2 - \delta r^2 - r^2 \delta \theta^2 - r^2 \sin^2 \theta \delta \phi^2, \quad (17-102)$$

where the space part is expressed in spherical polar co-ordinates (r, θ, ϕ) for the sake of convenience.

In the presence of the gravitating body, the metric gets altered because of the curvature of space-time introduced by it. Accordingly, one adopts the following ansatz by way of

generalization of the metric (17-102):

$$\delta s^2 = c^2 A(r) \delta t^2 - B(r) \delta r^2 - r^2 \delta \theta^2 - r^2 \sin^2 \theta \delta \phi^2, \quad (17-103)$$

where $A(r)$, $B(r)$ are functions to be determined by substitution in the Einstein equations. For the sake of clarity an index like μ will be assumed to take up values ranging over symbols 't', 'r', ' θ ', ' ϕ ' instead of '0', '1', '2', '3' respectively. Thus, as seen from the above expression for the squared space-time separation, the non-zero elements of the tensor $g_{\mu\nu}$ are

$$g_{tt} = A(r), \quad g_{rr} = B(r), \quad g_{\theta\theta} = r^2, \quad g_{\phi\phi} = r^2 \sin^2 \theta, \quad (17-104)$$

where the values of $g_{\theta\theta}$ and $g_{\phi\phi}$ are the same as in the case of Minkowski space-time while g_{tt} and g_{rr} are modified by the curvature caused by the gravitating body.

More specifically, one says that one can choose co-ordinates (the so-called Schwarzschild co-ordinates) t, r, θ, ϕ in terms of which the metric has the above form. I do not enter here into a discussion of the significance of these co-ordinates. However, these are related to the Local Lorentzian co-ordinates at every space-time point where the spatial location is outside the gravitating body.

With these assumed expressions for the metric tensor components, one can work out the Riemann tensor, the Ricci tensor, and the Ricci scalar, from which, finally, the Einstein tensor $G_{\mu\nu}$ can be worked out, where the latter comes out as a diagonal one, with G_{tt} , G_{rr} , $G_{\theta\theta}$, and $G_{\phi\phi}$ as its only non-vanishing components, involving the undetermined functions $A(r)$, $B(r)$.

As for the stress-energy tensor, all its components are zero for points exterior to the gravitating body:

$$T_{\mu\nu} = 0 \quad (\text{in free space; } \mu, \nu = 't', 'r', '\theta', '\phi'). \quad (17-105)$$

For points interior to the body, the stress-energy tensor at any space-time point depends on the density and pressure at that point under the condition of spherical

symmetry. In the static condition, the pressure counteracts the inward gravitational pull acting on any element of the body.

One can now substitute in the Einstein equations (17-99), and solve for $A(r)$, $B(r)$, making use of appropriate boundary conditions at $r \rightarrow \infty$, where the curvature of space-time has to reduce to zero value. We will be interested in the expressions for these two functions only for points exterior to the body ($r > R$) where, as mentioned above, all the stress tensor components vanish (the solutions at interior points are relatively more complex, and a complete picture is yet to emerge), so as to see how the latter causes space-time to be curved in the exterior region.

The result of this exercise gives us the following expressions for the functions $A(r)$, $B(r)$

$$A(r) = \left(1 - \frac{2GM}{c^2 r}\right), \quad B(r) = \frac{1}{1 - \frac{2GM}{c^2 r}}, \quad (17-106a)$$

and so the squared infinitesimal separation in curved space-time appears as

$$\delta s^2 = \left(1 - \frac{2GM}{c^2 r}\right) c^2 \delta t^2 - \frac{1}{1 - \frac{2GM}{c^2 r}} \delta r^2 - r^2 \delta \theta^2 - r^2 \sin^2 \theta \delta \phi^2. \quad (17-106b)$$

In these expressions, M is a parameter that turns out to be the Newtonian mass of the gravitating body, related to the density distribution ($\rho(r)$) as

$$M = 4\pi \int_0^R r^2 \rho(r) dr. \quad (17-106c)$$

This identification of the parameter M with the Newtonian mass follows when one considers the weak field limit of the Schwarzschild solution and compares the result with Newton's law of gravitation (refer to sec. 17.3.7.1).

While the Schwarzschild solution for the spherically symmetric static gravitating body looks quite simple, it leads to a number of important and interesting consequences, some of which I briefly indicate in the next section (sec. 17.3.7).

Though the above solution has been obtained on the assumption of a static source,

the same result obtains for a source with a variable mass distribution, for which the spherical symmetry is maintained at all times. Thus, a spherically symmetric source cannot produce a variable field at exterior points, regardless of the variation of the mass distribution within the source.

17.3.7 Schwarzschild solution: a few consequences

17.3.7.1 The Newtonian limit

The Schwarzschild solution correctly reproduces the Newtonian equation of a test particle (one that does not modify the gravitational field while having its own motion determined by the latter) in the field of the gravitating body in what may be called the ‘weak field’ limit (refer to sec. 17.3.3).

Generally speaking, the equation of motion of a test particle in the gravitational field is of the form (17-101), where the ‘force’ term F^μ actually arises by virtue of the curvature of space-time, and where the expression for F^μ involves the Christoffel symbols and the four-velocity components, calculated in terms of the parameter τ occurring in formula (17-101). Assuming that there is no other influence acting on the particle, i.e., that the particle is a ‘freely falling’ one, this equation of motion can be seen to imply the following interesting consequence: the trajectory of the particle between any two given space-time points on it is a *geodesic*, i.e., is one of minimum length as compared to the lengths of all other nearby paths connecting the same two points.

In other words, the effect of a gravitational field on a test particle is to *make it follow a geodesic path*. It is precisely this influence on the test particle that appears as a ‘force’ acting on the particle in the familiar Newtonian approach in a description of the motion of the particle.

For instance, in the weak field limit, where one can ignore terms of the second and higher degrees in $\frac{GM}{c^2 r}$, and where the speed of the particle is small compared to c ($|\frac{dx}{d\tau}| \ll c$), one obtains the result that the proper time τ equals the ‘time’ co-ordinate t occurring

in the Schwarzschild solution and, moreover, that the equation of motion reduces to

$$\frac{d^2 \mathbf{r}}{dt^2} = -\frac{GM}{r^2} \hat{e}_r, \quad (17-107)$$

where \hat{e}_r stands for the unit vector in the direction of increasing r , for constant values of θ, ϕ . In other words, the Schwarzschild solution correctly reproduces Newton's equation of motion in the weak field limit (with the parameter M identified as the Newtonian mass), and the Newtonian force of gravitation appears simply as the influence of the space-time curvature in this limit that causes the particle to follow a geodesic course.

The trajectory of the particle implied by the above Newtonian equation of motion is, in general, a *conic section* which, in the particular case of a bounded motion, is an *ellipse*, with the centre of the gravitating body as one of the two foci. However, when one considers the leading relativistic correction to the equation of motion implied by the Einstein equations, by keeping track of the second degree terms in $\frac{GM}{c^2 r}$, one finds that the trajectory is a *precessing ellipse*.

In the case of the sun as the gravitating body, the trajectories of planets are known to be ellipses, but astronomical observations indicate that these ellipses are also precessing slowly. The precession can be mostly accounted for in the Newtonian description itself in terms of the perturbation of the planetary motion caused by the other heavenly bodies in the solar system. In the case of the planet mercury, however, the observed perturbation cannot be fully accounted for in the Newtonian limit in terms of the effect of the other planets and satellites. Interestingly though, the discrepancy can be explained completely when one considers the relativistic correction to the equation of motion in the leading order.

17.3.7.2 Gravitational time dilatation and red shift

Let us consider an observer at rest relative to the spherically symmetric stationary gravitating body at a distance r from the centre ($r > R$), measuring time with a clock of her own, the time measured by the latter being say, T . Considering two events (say, E_1, E_2) close to each other occurring at the location of the observer, let the time interval be-

tween the two in the local Lorentz frame, i.e., the one measured by the observer's clock, be δT . Let this correspond to the increment δt in the first of the four Schwarzschild co-ordinates (the so-called time co-ordinate), the increment in each of the other three being zero. Then the invariant squared space-time separation between the events is given by

$$\delta s^2 = c^2 \delta T^2 = c^2 \left(1 - \frac{2GM}{c^2 r}\right) \delta t^2, \quad (17-108)$$

Let us now consider two similar events (E'_1, E'_2), now occurring at the location of a second observer, at rest relative to the gravitating body at a distance r' from the center, for which the increment in the first Schwarzschild co-ordinate is again δt , i.e., the same as for the previous two events. Let the time interval between these two events, as measured by the clock of this second observer (i.e., a clock that measures time in her local Lorentz frame) be $\delta T'$. The invariant squared space-time separation in this case is then

$$\delta s'^2 = c^2 \delta T'^2 = c^2 \left(1 - \frac{2GM}{c^2 r'}\right) \delta t^2. \quad (17-109)$$

Comparing the two equations, one obtains

$$\frac{\delta T}{\delta T'} = \sqrt{\frac{1 - \frac{2GM}{c^2 r}}{1 - \frac{2GM}{c^2 r'}}}. \quad (17-110)$$

Thus, if $r' > r$ then $\delta T' > \delta T$, i.e., the proper time interval gets dilated at the location with a relatively lower strength of the gravitational field - a phenomenon referred to as *gravitational time dilatation*. Recall that the proper time intervals $\delta T, \delta T'$ at distances r, r' do not correspond to a single pair of events, but to two pairs (E_1, E_2 , and E'_1, E'_2) for which the intervals measured in terms of the Schwarzschild time co-ordinate are the same.

A closely related phenomenon is the *gravitational red shift*, in which the frequency of a monochromatic wave gets decreased when observed at a point with a lower strength of the gravitational field (distance r') as compared with the frequency at a point with a relatively higher field strength (distance $r(< r')$).

Here the term ‘monochromatic wave’ needs explanation. If the Maxwell equations are written down in the Schwarzschild metric (17-106b), then it is found that these admit of solutions of the form $f(r)e^{-i\omega_0 t}$, i.e. ones that have a harmonic variation in the Schwarzschild time co-ordinate t , where ω_0 may be any arbitrarily specified frequency. These solutions of the Maxwell equations are precisely the monochromatic waves referred to above.

Consider two specified phases ψ , $\psi + \delta\psi$, close to each other of such a wave, and a situation where the two phases pass successively through the points P, P', both at rest relative to the gravitating body, at distances r , $r'(> r)$ from the centre. Let the passage of the two phases through P constitute the events E_1 , E_2 , while their passage through P' be identified as E'_1 , E'_2 . The intervals between the two pairs of events, measured in terms of the Schwarzschild time co-ordinate, are both equal to $\delta t = \frac{\delta\psi}{\omega_0}$.

Hence, the considerations leading to the relations (17-108), (17-109) apply, and one therefore arrives at the formula (17-110) relating the local Lorentz time intervals between the two pairs of events. These intervals, in turn, are related to the angular frequencies observed at the two points P, P' as

$$\omega = \frac{\delta\phi}{\delta T}, \quad \omega' = \frac{\delta\phi}{\delta T'}. \quad (17-111)$$

In other words, the frequencies of the monochromatic wave under consideration, as measured by the two observers at P, P' are related as

$$\frac{\omega'}{\omega} = \sqrt{\frac{1 - \frac{2GM}{c^2 r}}{1 - \frac{2GM}{c^2 r'}}}. \quad (17-112)$$

This tells us that the frequency of a monochromatic wave as observed at two different locations in a gravitational field differ from each other, the frequency being lower at the point where the gravitational field is relatively weaker. This, in summary, is the phenomenon of the *gravitational red shift*.

17.3.7.3 Black holes

A look at the Schwarzschild metric (17-106b) tells us that it has a singularity at $r = \frac{2GM}{c^2}$ ($= r_S$, say) since the coefficient of δr^2 blows up at this value of r . For a large class of heavenly bodies such as the sun and the stars, this corresponds to points deep in the interior, and is not of great contemporary interest (recall that the solution (17-106b) holds in the exterior region; the interior solution involves more complex consideration). For a sufficiently massive gravitating body, however, the above value of r corresponds to a surface lying in the exterior region and poses the question as to what it signifies.

In the first place, the singularity at $r = r_S$ is not a 'genuine' one, and is actually an undesirable feature of the choice of co-ordinates. For instance, the modulus of determinant of the metric tensor, as also the Riemann tensor remains well behaved at this value of r . What is more, the time of free fall of a particle from any chosen point (with $r > r_S$) down to $r = r_S$ is seen to blow up when worked out in terms of the Schwarzschild time t , but works out to a finite value when calculated in terms of the proper time of the particle.

Alternative co-ordinate systems can be defined in terms of which, the singularity in the metric disappears. However, even in these co-ordinates, one finds that a set of extraordinary features are associated with the surface $r = r_S$. For instance, once a particle enters into the interior of this surface, it continues to go straight down to $r = 0$ which constitutes a genuine space-time singularity of the gravitational field (i.e., it cannot be got rid of by any alternative choice of co-ordinates). Additionally, any signal emitted from such an interior point (interior, that is, to the surface $r = r_S$, and not to the boundary surface of the gravitating body) does not reach distant observers ($r \rightarrow \infty$), i.e., the future light cone lies within the interior, and does not communicate with exterior points. The surface $r = r_S$ is therefore referred to as the *event horizon*.

In other words, no events occurring within the surface $r = r_S$ (and hence within the boundary of the gravitating body itself, which lies completely within the surface) can be recorded or observed by an external observer - the body in question appears as a *black hole*.

Numerous black holes have been detected in the skies by indirect means, lending support to the general theory of relativity and, in particular, to the Schwarzschild solution. Black holes are relevant in the context of *life histories of stars* since, a star, in the course of *gravitational collapse* (i.e., the process of contraction caused by the fact that the burning of nuclear fuels does not generate enough outward pressure so as to counterbalance the inward gravitational pull) can turn into a black hole.

17.3.8 The general theory of relativity: the classical and the quantum

The general theory of relativity is a *classical* theory of gravitation interacting with particles (as also with photons) where the gravitational field bends the trajectories of particles along geodesics while, at the same time, swarms of particles generate a curvature of space-time. As a classical theory, it works well at *large scales* where quantum effects are not of relevance.

At much smaller scales, on the other hand, quantum uncertainty effects must inevitably acquire greater importance, making the classical theory of general relativity inadequate.

In section 5.6, I gave you a few sketchy remarks on how the Newtonian view of gravitation gets broadened into the viewpoint of the general theory of relativity. In the sections on the general theory in the present chapter, I have meant to elaborate on those remarks by way of explaining a few basic ideas that lead to such a broadening of the viewpoint.

At the same time, I pointed out in sec. 5.6 that the general theory of relativity, in remaining confined within the limits set by the *classical* point of view, itself needs to have fresh ideas injected into it, whereby it can fit in a quantum field theoretic framework. On the face of it, the necessity for such fresh ideas may appear to be of nothing more than an academic interest since gravitation is relevant in describing large scale phenomena while quantum field theory is concerned with phenomena at the smallest scales of length, mass and time.

Such a neat compartmentalization of the scopes of the two theories, however, breaks

down in the case of phenomena where the space-time geometry assumes relevance even at smallest conceivable scales. For instance, matter and radiation inside a black hole may be pulled inwards by gravity so as to approach arbitrarily close to the singularity at the origin. The classical framework of the general theory of relativity becomes inadequate for the explanation of the likely range of phenomena in such a situation involving an enormous concentration of mass-energy within an inconceivably small spatial range, because conceptual barriers of a fundamental nature come up in any attempt at such an explanation. These, precisely, are the barriers relating to the *Planck scales* of length, mass and time, which come up because the Planck constant becomes relevant in the explanation of gravitational phenomena, i.e., *quantum fluctuations* of the gravitational field assume an overriding relevance.

For instance, the Planck length ($l_P = \sqrt{\frac{\hbar G}{c^3}}$) comes up when the Schwarzschild radius (r_S , see sec. 17.3.7.3) of a gravitating particle overlaps its *Compton wavelength* (the wavelength at which Planck's expression for the quantum of energy equals the mass-energy). The need for a consistent quantum theory of gravitation comes up as this scale of length is crossed (the Planck scales of mass and time are similarly arrived at).

The singularity at $r = 0$ in the problem of the spherically symmetric gravitating body is therefore only an ideal one since the Planck scale of length is crossed before the singularity is reached, and completely new sets of phenomena, requiring completely new ideas for their explanation, are likely to emerge. Such new ideas are also necessary to analyse the likely range of phenomena in the very initial stages of the *big bang* (within the Planck scale of time $t_P = \sqrt{\frac{\hbar G}{c^5}}$).

For an introduction to the general theory of relativity you can look up [14]. There are, however, many other renowned texts on the general theory.

Chapter 18

Atoms, Nuclei, and Molecules

18.1 Introduction

All matter is composed of *molecules*. The molecules of a pure substance - either an element or a compound - are the smallest structural units that possess the chemical properties of the substance. Molecules, in turn, are made up of *atoms*. For instance, a molecule of water, H_2O , is made of two atoms of hydrogen and one atom of oxygen. Some elements like helium are monatomic, which means that a helium molecule is the same as a helium atom. Different elements differ in the structure of their atoms, and it is the atom that determines the manner an element combines with others to form compounds.

The search for elementary or ultimate constituents of matter and for an explanation as to how these constituents are held together so as to form the various materials we see around us, is an age-old one. And it relates to the question as to how the fascinating variety and complexity of the properties of the various different materials can be accounted for in terms of the properties of their constituents. However, it is only in comparatively recent times that this search has been based on well-thought-out experimental investigations, without pre-conceived notions playing a dominant role in it.

The present-day theory of the atom was initiated through a series of theoretical and

experimental investigations in the early twentieth century. To begin with, a number of investigations had been pointing towards inconsistencies in the classical view of nature, which led to the introduction of novel ideas, thereby inaugurating the era of *quantum theory* (see chapter 16). These ideas were made use of, notably by Bohr, in working out a successful explanation of the spectrum of the hydrogen atom.

Bohr's approach was further pursued by Sommerfeld, and this theoretical development was supplemented by experimental investigations of Rutherford and co-workers relating to the atomic nucleus, yielding the *Bohr-Rutherford planetary model of the atom*. In the next section I am going to outline the basic features of this model without going into an account of the long series of theoretical and experimental investigations that led the pioneers like Bohr, Sommerfeld and Rutherford, and their successors, to piece together the present-day picture of the atom. This picture includes a number of basic facts relating to the atomic *nucleus* as also to states of electrons around the nucleus. You will find that some of the material presented below will make better sense when read with parts of chapter 16.

18.2 The atomic nucleus: atomic volume and mass

The atom can be thought of as something like a miniature solar system, with the nucleus at the center and with electrons surrounding the nucleus in a number of *shells*. The nucleus occupies a tiny, almost insignificant, fraction of the volume of the atom while almost the entire *mass* of the atom is concentrated in the nucleus. The typical linear dimension of the atom is of the order of 10^{-10}m . By contrast, the nucleus is typically of the order of 10^{-15}m in linear dimension, which makes the nuclear volume 10^{-15} times the atomic volume in order of magnitude.

The nucleus is made up of *protons and neutrons*, collectively termed *nucleons*. The nucleons are all of approximately the same mass (nearly $1.66 \times 10^{-27}\text{ kg}$), with a neutron being slightly heavier than a proton. By comparison, an electron is much, much lighter, with a mass of approximately $9.1 \times 10^{-31}\text{ kg}$.

The proton and the electron carry opposite charges - respectively positive and negative - of the same magnitude (nearly 1.6×10^{-19} C), while the neutron carries no charge. The number of protons in the atomic nucleus of an element, which is also equal to the total number of electrons in the various shells surrounding the nucleus, is termed the *atomic number* of the element under consideration and is commonly denoted by the symbol Z . Thus, denoting the magnitude of the electronic charge by e , the charge in the nucleus equals

$$Q_{\text{nucleus}} = +Ze. \quad (18-1)$$

A neutral atom consists of electrons, held in *orbitals* (see below, sec. 18.3.2, as also sec. 16.10.3) around the nucleus, the number of these extra-nuclear electrons being the same as the number of protons (Z) inside the nucleus. The total charge of these electrons is then $-Ze$. By contrast, an *ion* is a charged atom where one or more electrons from the orbitals have been removed (negatively charged ions, with electrons added to the atoms, are also possible).

The total number of nucleons in the nucleus is termed the *mass number* and is commonly denoted by the symbol A . The number of neutrons is then given by $A - Z$, and the nuclear mass can then be expressed as

$$M = (A - Z)m_n + Zm_p, \quad (18-2a)$$

where m_n and m_p stand for the mass of the neutron and proton respectively. This is an approximate expression since it gives the mass of the neutrons and the protons *separated from one another*. The nucleus, on the other hand, is a *bound* structure where the neutrons and the protons are held together by certain binding forces (more on this later), and it requires some energy to get them separated from one another. By the principle of *mass-energy equivalence* (refer to section 17.2.8.6), this shows up as a deficit of the nuclear mass relative to the mass of the separated nucleons. This deficit, known as the *nuclear binding energy* (see sec. 18.8.4), when expressed in energy units, is important in the context of the structure and properties of the nucleus, but is not

of direct relevance while talking of the characteristic properties of the atom as a whole since the latter are determined principally by the extra-nuclear electrons.

The expression (18-2a), an approximation in itself, can be further approximated if we recall that the neutron and proton masses are almost equal:

$$M \approx Am_p, \quad (18-2b)$$

and, ignoring the mass of the electrons (much smaller than m_n or m_p), and the binding energy of the electrons in the atom, one can also take it to be an approximate expression for the mass of the atom.

While this gives an estimate for the atomic mass, a similar estimate for the atomic volume is more difficult to come by. The atomic volume can be defined as the radius of the last occupied electronic shell (see below, sec. 18.4.1), though the radius of a shell is itself not a precisely defined quantity. As I have already told you, the nucleus occupies an insignificant part of volume of the atom, the linear dimension of the latter being of the order of 10^{-10} m.

18.3 Single-electron states

Imagine that you have an atomic nucleus at your disposal and you are building up the atom by bringing in electrons in succession, adding one electron after another.

As you bring in the first electron, it feels the Coulomb attraction due to the positive charge of the nucleus, the magnitude of force being given by

$$F = \frac{1}{4\pi\epsilon_0} \frac{Ze^2}{r^2}, \quad (18-3)$$

which is similar to the force on the single electron in the hydrogen atom (sec. 16.10), differing only in the factor Z arising due to the number of protons in the nucleus being Z rather than unity, the latter corresponding to the single proton in the hydrogen atom.

One can then go on step by step with the derivations in the Bohr theory of the hydrogen atom, keeping in mind this distinguishing feature relating to the nuclear charge, and come up with a number of useful conclusions.

It has to be mentioned at this point that the *effective* charge felt by an electron, differs from the nuclear charge Z (in units of the electronic charge) as the electrons are added one after another. In addition, this effective charge depends on the effective distance of the electron from the nucleus. This will be an issue of interest in what follows.

As an electron is added, it forms a bound system with the nucleus, and can be caught in any one of the possible stationary states, each characterized by a quantum number n and a value of the energy, given by

$$E_n = -chRZ_{\text{eff}}^2 \frac{1}{n^2}, \quad (18-4a)$$

where Z_{eff} stands for the effective charge mentioned above, and R for the Rydberg constant defined in sec. 16.10.2:

$$R = \frac{me^4}{8\epsilon_0^2 h^3 c}, \quad (18-4b)$$

whose value is

$$R = 1.1 \times 10^7 \text{ m}^{-1} \text{ (approx)}. \quad (18-4c)$$

In this theory, which is not yet the full quantum theory of the atom, one can talk of an electron orbit with quantum number n , an estimate for the radius of the orbit being obtained from

$$r_n = \frac{\epsilon_0 h^2}{\pi m Z_{\text{eff}} e^2} n^2. \quad (18-5)$$

I will come back to this story of the atom being built up by the successive addition of electrons later, after I tell you something more of the possible states of a single electron

in the Coulomb field of the nucleus.

I want you to get the picture clear at this point. We are now talking of a many-electron atom with a nuclear charge Z but, at the same time, are making use of a simplified scheme of description in which the electrons are considered to be moving independently of one another. In reality, this is not *legal* since, in reality, the electrons interact with one another. In considering the possible states a single electron, we assume for the time being that these are similar to those in an atom with one single electron, but with the effect of the other electrons taken into account by replacing the nuclear charge Z with some effective charge Z_{eff} . It turns out that this simple assumption is not an arbitrary one since it can be interpreted in terms of the concept of *screening* to be discussed later. With this modification introduced in respect of the nuclear charge, the *single-electron* states will refer to the states of a single electron in a many-electron atom, assumed to be moving independently of the other electrons in the same atom. As we will see, the simplification has a price tag attached to it since Z_{eff} is not an easy thing to determine quantitatively and, moreover, depends on the state of the electron under consideration.

18.3.1 Single-electron states: elliptic orbits and degeneracy

The single-electron states in a many-electron atom are more complex to describe than those in an atom with a single electron. In the present section we consider, for the sake of future reference, the possible states in an atom with a single electron, pointing out a number of modifications and generalizations over the Bohr theory.

Bohr's theory of the hydrogen atom (see sec. 16.10), is incomplete on two counts. First, it refers to electron orbits, which presupposes that the electron has a precisely defined position at every instant of time while, in reality, the position is more truthfully described as a random variable (see sec. 16.2). And secondly, while talking of electron orbits, it refers to only the circular ones, disregarding the possibility of elliptic orbits. In fact, elliptic orbits arise as the general solution for bounded motion of a particle in an attractive inverse square field of force (see sec. 5.5 in connection with planetary motion), and circular orbits then constitute only special cases of these more general solutions.

Sommerfeld improved upon Bohr's theory for a single-electron atom by including elliptic orbits into consideration and obtained the important result that the quantum number n gives only a partial description of the possible stationary states of the electron. He found that, for any given value of n , the electron can, in general, be in one of *several* possible stationary states, corresponding to a *number of possible elliptic orbits*, this number being precisely n . For each of these n number of orbits, the electron has a precise value of the angular momentum about the nucleus (in contrast to the position co-ordinates of the electron that correspond to random variables) which differs from the value of the angular momentum for any other of the possible set of elliptic orbits. Related to the different possible values of the angular momentum, the orbits differ from one another in the value of the *eccentricity* of the ellipse.

Analogous to the quantum number n that determines the energy of the electron in Bohr's theory, one can define another integral quantum number l which determines the angular momentum of the electron in an elliptic orbit, thereby telling us which of the n number of possible elliptic orbits the electron describes. For a given value of n , which we now refer to as the *principal* quantum number, this second quantum number l , termed the *angular momentum quantum number*, can assume integral values ranging from 0 to $n - 1$, making it a total of n different orbits. Of these, *the one with the maximum possible value of l , i.e., the one with $l = n - 1$, corresponds to the circular orbit of Bohr's theory.*

However, I have to qualify my statements here. In reality, the quantum number introduced by Sommerfeld differed in some respects from the quantum number l I am talking of since he did not base his considerations fully on quantum principles. The new quantum number he introduced led to essentially correct conclusions, though. While the quantum number l is mentioned here in the context of Sommerfeld's work, it is accounted for by quantum considerations of later day origin.

As an example, the two possible orbits for $n = 2$ are shown schematically in figure 18-1. The possible values of l here are $l = 0$ and $l = 1$, of which the former corresponds to an elliptic orbit while the latter to the circular Bohr orbit. Incidentally, there is just one single orbit ($l = 0$) for $n = 1$, this being a circular orbit.

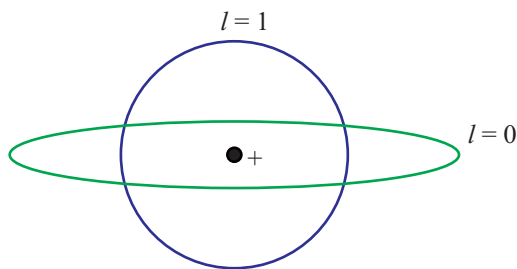


Figure 18-1: A circular and an elliptic orbit, with $l = 1$ and $l = 0$ respectively, of an electron around a positively charged nucleus for principal quantum number $n = 2$ (schematic); Bohr's theory allows for only the circular orbit while Sommerfeld's extension of Bohr's theory allows for both.

Thus, in this extended version of Bohr's theory, a stationary state of the electron is determined not by one but by *two* quantum numbers n and l . One more finding of the theory is that the energy of the electron is determined solely by n and is given by the *same* expression (equation (18-4a)) that Bohr's theory gives. All the n number of states for principal quantum number n are then characterized by the same energy but by different values of the angular momenta. Since Bohr's theory did not accommodate the second quantum number l , there was no question of the various possible states with the same energy, and the only possible state for a given n was characterized by a single value of the angular momentum, given by Bohr's basic equation (16-36). What is important to note is that, similar to the energy, the possible angular momentum values of the electron also form a discrete set or, in other words, the angular momentum is *quantized*.

Sommerfeld's theory also indicated that the energy of a stationary state had a weak dependence on the second quantum number l , owing to relativistic effects relating to the motion of the electron around the nucleus, but this I will not elaborate upon in the present introductory exposition.

This phenomenon of the existence of several different states with the same value of energy (disregarding the weak dependence of the energy on the quantum number l) is known as *degeneracy* in quantum theory.

18.3.2 Single-electron states: space quantization

While the new quantum number l gives a more complete description of the stationary state of the single electron by specifying the angular momentum and the shape of the elliptic orbit corresponding to the state, it still does not specify the state uniquely. In other words, there may still be more than one distinct stationary states with identical values of n and l . These various stationary states are distinguished from one another by the *orientation in space* of the plane of the orbit corresponding to the state. A crucial result is that *this orientation is also quantized*. In other words, choosing any specific direction in space (let us call it the z-axis), the angle between this direction and the plane of the orbit cannot be just anything - it has to have a value belonging to a *discrete* set. These possible orientations of the orbit in space are distinguished from one another by the value of a *third* quantum number referred to as the *azimuthal*, or *magnetic* quantum number and commonly denoted by the symbol m_l . For any given value of l (recall that this can be any integral value ranging from 0 to $n - 1$), m_l can have any integer value from $-l$ to $+l$, i.e., the number of possible orientations in space is $2l + 1$.

Once again, Sommerfeld's conclusions, having been based on partly classical considerations, differed in details from those relating to space quantization (see below) that follow from quantum considerations.

Figure 18-2 depicts schematically the three possible orientations of an orbit with $l = 1$, for which the azimuthal quantum number m_l can have values -1 , 0 , and $+1$. In this figure I have drawn an imagined orbit with dotted lines oriented in between the orbits with $m_l = 0$ and $m_l = 1$. The all-important result is that *quantum theory rules out such an orbit* because only the three orientations shown with solid lines are allowed by the fundamental principles of the theory. This discreteness in the orientations of the orbits corresponding to possible stationary states is termed *space quantization*. One consequence of space quantization is that the possible values of the component of the angular momentum along the z-axis (the vertically upward direction in fig. 18-2; any other direction could also be chosen as the reference direction, with respect to which all the orientations would then have to be referred to) form a discrete set. In other words,

not only the angular momentum but its z -component is also quantized.

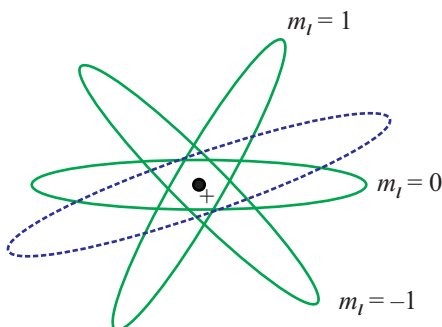


Figure 18-2: Three possible orientations of orbit for $l = 1$, corresponding respectively to $m_l = -1, 0$ and 1 ; the hypothetical orbit shown with a dotted line is ruled out; the orientations are referred to in terms of a chosen direction which in this case is the vertically upward direction - the z -axis of a Cartesian co-ordinate system.

In so far as the stationary state of an electron with given values of the quantum numbers n, l, m_l can be associated with a classical orbit of the electron, it is sometimes referred to as an *orbital*. However, this term is also used in a more general context in referring to single-electron states in atoms or molecules even when no such correspondence with classical orbits is implied. More specifically, the term orbital refers to a quantum state without regard to the *intrinsic* state (see sec. 18.3.3) of the electron.

18.3.3 Single-electron states: electron spin

As I have told you, the make-shift theory of Bohr and Sommerfeld I have been outlining here talks in terms of classical concepts while at the same time making clever use of a number of novel ideas and yielding a number of important results. This theory was superseded by the full quantum theory of the atom where the classical concepts are not referred to, but a number of results derived in this new theory *agreed* with the basic results of the earlier Bohr-Sommerfeld theory outlined above. As I have mentioned, I have chosen to deviate somewhat from Sommerfeld's version that improved upon Bohr's theory, by introducing quantum numbers differing from Sommerfeld's, in line with what later day quantum theory brought to light.

However, the Bohr-sommerfeld theory still fell short of providing a *complete* description of the single-electron stationary states because it held no clue to what was subsequently recognized as a *specifically quantum feature* of the electron (and of other elementary particles in nature), with no classical interpretation, namely its *spin*.

Though one cannot talk of the spin in classical terms without encountering inconsistencies and contradictions, an *analogy* can be given by referring to the diurnal rotation of the earth about its axis. While the quantum numbers l and m_l relate to the *orbital* motion of the electron that can be compared with the annual motion of the earth around the sun, the spin compares to the diurnal motion. One more similarity in this analogy is that the spin is responsible for an *angular momentum* of the electron in addition to that arising due to the orbital motion.

In reality, the spin is an *internal* feature of the electron, quite distinct from features relating to orbital motion. The existence of spin was first made evident in a famous experiment by Stern and Gerlach. Similar to the fact that the orbital angular momentum and its z-component are determined by the quantum numbers l and m_l , the angular momentum due to spin, and the z-component of this angular momentum, are determined by two new quantum numbers, commonly denoted by s and m_s . However, an elementary particle like an electron or a proton is characterized by a *fixed* value of the *spin quantum number* s , which can be either a non-negative integer or a *half-integer*. For an electron (as also for a proton) this fixed value is $s = \frac{1}{2}$. For any given value of s , the *magnetic spin quantum number* m_s can take values ranging from $-s$ to $+s$, where two successive values differ from each other by unity. Thus, for an electron, there can be *two internal states* corresponding to $m_s = -\frac{1}{2}$ and $m_s = +\frac{1}{2}$.

The term 'magnetic' is used while referring to m_l, m_s , since these quantum numbers assume relevance as one attempts to describe the magnetic properties of atoms.

While the quantum numbers m_l and m_s are needed for a complete specification of the stationary state of the electron, they do not have any bearing on the *energy* of the stationary state, which continues to be given by the Bohr expression (18-4a), in which

an effective nuclear charge Z_{eff} (in units of the electronic charge) has been used in the place of Z . In this they resemble the angular momentum quantum number l , but only up to a point. As mentioned above, a deeper analysis shows that the energy does have a weak dependence on l due to relativistic effects, i.e., states with the same n but different values of l are not quite degenerate - they have slightly different energy values. However, there remains the degeneracy with respect to m_l and m_s , i.e., states with the same n and l but different values of m_l and m_s have the same energy, i.e., such states are, in the true sense, degenerate.

The degeneracy with respect to m_l, m_s disappears when the atom is placed in a magnetic field. However, precise statements in this regard require a more careful description and analysis of atomic states in the presence of magnetic fields.

18.3.4 Single-electron states: summary and notation

It is now time to sum up all I have said above relating to stationary states of a single electron in the attractive Coulomb field of a nucleus. In such a field the electron can be caught in one of several *bound stationary states* that can be catalogued by referring to possible classical orbits of the electron. According to Bohr-Sommerfeld quantization rules, only some of these orbits, forming a discrete set, can correspond to stationary states of the electron. A stationary state is completely specified by three orbital quantum numbers (n, l, m_l) and an internal quantum number (m_s), the latter relating to the spin of the electron. The possible values of these quantum numbers are: $n = 1, 2, 3, \dots$; $l = 0, 1, \dots, n - 1$ (for a given n); $m_l = -l, -l + 1, \dots, l - 1, l$ (for a given l); and $m_s = -\frac{1}{2}, +\frac{1}{2}$ (for the spin quantum number $s = \frac{1}{2}$ of the electron).

The *energy* of the stationary state depends, in the main, on the principal quantum number n , and increases with increasing value of n . It also depends on the angular momentum quantum number l , but to a much lesser extent - the energy increases by a small amount as l is made to increase. This gives us a clue as to how to arrange the various stationary states in order of increasing energy. Thus, for instance the state with $n = 2, l = 1$ has a (slightly) higher energy than the one with $n = 2, l = 0$, while the energy

for $n = 3, l = 0$ is higher than both.

There is a convention for using symbols for the single-electron states with various possible values of n and l using the numerical value of n and certain letters of the English alphabet to represent the values of l . For instance, the letters 's', 'p', 'd', and 'f' are used to represent $l = 0, 1, 2, 3$ respectively, while higher values of l are represented by the letters 'g', 'h', etc. Thus, the state $n = 2, l = 0$ is represented by the symbol 2s and the state with $n = 3, l = 2$ by 3d according to this convention.

How many distinct states can there be for a given value of n ? With n fixed, the different possible states correspond to distinct combinations of the remaining three quantum numbers l , m_l , and m_s . According to the rules outlined above, the number of such combinations turns out to be $2n^2$. All these $2n^2$ number of states have almost the same energy, though the ones with higher l values have slightly higher energies compared to others. Finally, the number of states with given values of n and l works out to $2(2l + 1)$, all of which are degenerate, i.e., have the same value of energy (or, nearly the same value).

All these statements are corroborated by the full quantum theory of the electron in an attractive Coulomb field, although this theory is based on a point of view quite radically distinct from that of the Bohr-Sommerfeld theory. The full quantum theory does not refer to electron orbits and works out the stationary states as solutions of what is termed the *Schrödinger equation*. Yet, remarkably, the basic findings of the Bohr-sommerfeld theory *augmented with the concept of the electron spin* are found to remain valid in the more advanced quantum theory as well.

18.4 Building up the atom

18.4.1 Electronic configuration and electron shells

With all the necessary preliminaries having now been cleared, let me return to the story I started with in this section, namely that of building up of an atom by an imagined

process of successive addition of electrons to the nucleus. As the first electron is added, it can be caught in any one of the possible stationary states but if that state happens to be higher in the energy scale compared with the lowest energy state (or the 'ground state', as it is called) namely a 1s state (recall that there can be two degenerate 1s states) then it can emit a photon to make a transition to a state with a lower energy (see sec. 16.10.4) and the process can continue till it reaches the ground state. In other words the stable state for the first electron to be in is a 1s state. I may as well point out here that while it is possible for an electron to make a transition from a higher energy state to one with a lower energy (provided that such a lower energy state is 'available', see below) by the emission of photons, the reverse process of moving *up* the energy scale is not possible unless photons are made available to the atom for it to *absorb* one or more of those. This means that the electron can get *excited* to a higher energy state only if an *external* electromagnetic field, or some other source of energy, helps it to.

All right, then, with the first electron lodged in a stable 1s single-electron state, what happens when the *next* electron is brought in? The answer is simple in principle: it too will get caught in the lowest energy state *available* to it. What needs explaining here is the word 'available'. And *that* brings in yet another celebrated principle in quantum theory:

Pauli's exclusion principle: A single-electron state, specified by quantum numbers n, l, m_l, m_s , cannot accommodate more than one electron.

This simple-looking principle explains a lot of things. First, recall that there are *two* 1s single-electron states, both of which are states with the lowest possible energy, and, of these, *one* has already been occupied by the first electron. According to Pauli's principle, *this* 1s state cannot accommodate the second electron, and the lowest energy state now *available* to the second electron is the *other* 1s state. Thus, with the introduction of the second electron, both the 1s states are now 'filled up' - Pauli's principle does not allow any other electron to come in and take shelter in any of these two states.

Incidentally, a principle of great relevance in quantum theory is that of *indistinguishability*.

bility: electrons are fundamentally indistinguishable from one another, as a result of which it is not quite meaningful to refer to one of a pair of electrons as the *first* one and to the other as the *second*. One can say that two electrons are lodged in the two $1s$ states without, however, trying to specify *which* electron is in which of the two state (the *states* are not indistinguishable, but the electrons are). Interestingly, Pauli's exclusion principle comes out as a corollary to the principle of indistinguishability (refer to sec. 18.5.2.1).

Now imagine a *third* electron to be brought in (once again, this way of describing things is in violation of the indistinguishability principle; I adopt it only for the sake of better intelligibility). It too will be caught in the lowest available single-electron state. But since it is denied berth to any of the two $1s$ states, the lowest energy state available to it is $2s$ (check this out by going back to the results I summarized above). There are two such states, of which one will be occupied by this third electron while, if a fourth electron is introduced, it will be accommodated in the other $2s$ state. The fact that, after the introduction of the fourth electron, two are lodged in $1s$ states and two others in $2s$ states is expressed symbolically by saying that the *electronic configuration* of the resulting system is $1s^2 2s^2$.

The story continues. Every time an electron is brought into the attractive Coulomb field it 'seeks out' the lowest energy single-electron state available to it, these single-electron states being filled up successively from bottom upwards in the energy scale according to Pauli's principle, determining the electronic configuration of the bound system at every stage of the process. For instance, with *five* electrons forming a bound system with the nucleus, the electronic configuration would be $1s^2 2s^2 2p^1$.

The sequence of filling up of the single-electron states is, however, not a simple one, where states with a given n are filled up in a sequence of increasing l values and, once all the possible l -states (with various possible m_l and m_s values) are filled up, the next higher n -state (the next 'shell' see below) starts being populated. In reality, this simple rule needs to be modified into one with a broader scope where the ordering of energy values in terms of n and l is somewhat different, as explained below in sec. 18.5.2.2.

Even so, the modified rule is not one of general validity since one finds quite a number of exceptions to it as one examines elements with relatively large atomic numbers in the periodic table. The principle relating to the filling up of successive energy levels ('shells' and 'subshells', see below) is thus nothing more than a convenient simplification - not a logically compelling one - of what is actually a complex reality relating to the structure of atoms.

The actual sequence of energy values of the single-electron states is determined by the *interactions* among the various electrons in an atom. A simple way to account for such interactions is to resort to the idea of *screening* of the nuclear charge (sec. 18.5.1). According to this idea, successive electrons brought in in the process of building up of the atom experience the Coulomb attraction of the nucleus with a *decreased* effective charge Z_{eff} . In other words, when the electrons in the atom are considered together, the concept of the single-electron energies gets altered, with the value of Z_{eff} being different for the various states. Another aspect of the complexity of the many-electron atom is revealed by the fact that the energy of a single-electron state acquires an l -dependence, that is more pronounced than the weak l -dependence, mentioned above, arising in the case of an atom with a single electron.

Note that there are two things at work here - the ordering of the single-electron states with increasing energy, and Pauli's principle. In the case of relatively light atoms, the electrons tend to fill up all states with the lowest available value of n , after which the states with the next higher value of n are approached. For a given n , states are filled up in order of increasing l . These facts are expressed by saying that all the states with a given n belong to a *shell* and, within a shell, all states with a given l form a *subshell*, the subshells and shells being filled up sequentially.

The energies of successive shells differ by a considerably larger extent compared to the difference between successive subshells within a shell. The successive shells with $n = 1, 2, 3, \dots$, are referred to respectively as the K, L, M, ... shells, while subshells with $l = 0, 1, 2, \dots$ are referred to as respectively the s, p, d, ... subshells. As all the single-electron states in a shell or subshell are filled up, one says that, that shell or subshell

is *closed* or completed.

However, as indicated in the notes above, the principle of filling up of successive subshells and shells cannot be described, generally speaking, in such simple terms, especially for the heavier atoms, since this requires more complex considerations (refer to sec. 18.5.2.2 where a more general principle of filling up of single-electron states in shells and subshells is introduced).

18.4.2 Electronic configurations and the periodic table

This approach of understanding the structure of the atom by imagining electrons being added one after another is known as the *aufbau* (which means ‘building up’) principle, first outlined by Niels Bohr. Consider, for instance, a helium atom ($Z = 2$) being built up this way. The neutral helium atom has two electrons around its nucleus, and the aufbau principle tells us that its electronic configuration would be just $1s^2$ (by contrast, an *ion* is a charged atom, with an excess or a deficit of electrons compared to the number of protons, and a He^+ ion, for instance, with just one electron in it would have the *hydrogen-like* configuration $1s^1$).

This means that the two electrons in the neutral helium atom form a closed shell, namely the first, K shell. The lithium atom, on the other hand, contains three electrons, of which two are lodged in a closed K shell while the third resides in the next, L shell, corresponding to the configuration $1s^2 2s^1$. The neon atom with $Z = 10$ is made up of closed K and L shells, while the sodium atom with $Z = 11$, having an electronic configuration $1s^2 2s^2 2p^6 3s^1$ has closed K and L shells, together with one electron in the M shell.

Indeed, the *periodic table* is built up, in the main, on the basis of the aufbau principle relating to the electronic configurations of atoms. It predicts that there has to be a certain periodicity in the properties of elements based on the order of their position in the periodic table, the latter being determined by their atomic numbers.

As we have seen, the electronic configuration of an atom consists, in general, of one

or more closed shells and subshells together with one or more electrons in a partially filled subshell. The number of electrons in the last unfilled subshell are termed *valence electrons*. Electrons can be added or removed from this subshell comparatively easily, with comparatively little energy change, and it is this that determines the chemical reactivity and a number of associated properties of the element under consideration.

The elements like helium and neon, having a closed-shell electronic configuration are the least reactive, and are termed the inert gases. One then finds a periodic recurrence in the number of electrons in the unfilled shell or subshell for elements occurring in between the inert gases in the periodic table. For instance, lithium and sodium are located immediately after helium and neon in the periodic table, with one electron each in the last, unfilled s-subshell, and have similar properties, being alkaline metals both.

18.5 The atom as a whole

18.5.1 Screening of the nuclear charge

As mentioned above, the atom is not made up of electrons bound by the nucleus *independently* of one another. These electrons have a repulsive Coulomb interaction among themselves and, as a result, the energies of the available single-electron states at any given stage of the building up process depend in a non-trivial manner on the electrons that have already been assembled.

What one has to do in understanding the behavior of an atom is to describe the possible stationary states of the *atom as a whole*, taking into account the interaction among the electrons.

This is a complex problem in quantum theory and the last word is yet to be said. Still, one can make a number of statements based on relatively simple considerations and a few empirical observations. The justification of these statements from the full quantum theory needs a deeper analysis, and is a problem of quite considerable complexity, which I will only briefly refer to in a later section (see sec. 18.5.2). Most of these statements

relate to the problem of determining the *ground state*, i.e., the state with the lowest energy, of the atom and, in some instances, the frequencies characterizing transitions from relatively low-lying *excited* states to the ground state or to other excited states. I address first the ground state problem and begin with a simple example, namely, that of the *helium atom*. The problem of describing low-lying excited states in a few simple instances will also be addressed. A class of problems involving transitions will then be taken up in connection with *X-ray spectra*, which again requires relatively simple considerations.

As we have seen, the electronic configuration of helium is $1s^2$. But this by itself does not help one determine the ground state energy of the atom or the wave function of the atom in the ground state. In order to work these out, one has to go to a deeper level, taking the mutual interaction of the electrons into consideration. The ground state energy of the atom that comes out of this exercise happens to have a nice interpretation in terms of what is referred to as *screening*.

One finds that this energy can be expressed as the sum of two terms representing the energies of the two electrons in a Coulomb field, the two energies being equal because of the indistinguishability of the two electrons. However, now the Coulomb field is *not* that of a nuclear charge $2e$, *but one of charge $1.7e$* (approx). This means that the behavior of any one of the two electrons is not determined solely by the nuclear charge $2e$ but by the other electron as well, where the influence of the latter can be effectively described as a *reduction of the nuclear charge* by an amount $0.3e$. In reality, an atomic electron in quantum theory is somewhat like a smeared cloud of negative charge around the nucleus and a second electron moving around the nucleus through this cloud finds part of the positive nuclear charge neutralized or screened by this cloud (see figure 18-3).

More generally, in an atom with several electrons and with nuclear charge Z (in units of the electronic charge), the effect of inter-electron interaction on any one of the electrons can be described in an approximate sense as a partial neutralization or screening of the nuclear charge experienced by it. The energy of the atom can then be approximated as

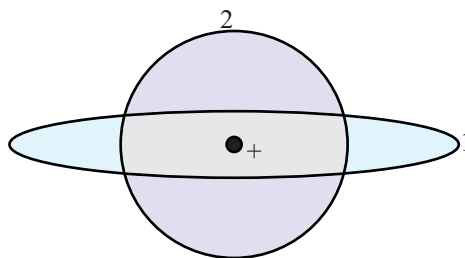


Figure 18-3: Illustrating the concept of screening of the nuclear charge by one electron with reference to another; two single-electron orbits are shown; the orbit marked 1, is in reality, a smeared charge-cloud (indicated by the shading; while a uniform shading is shown in the figure, in reality the cloud is relatively thick in some regions and thinner in some others) through which the orbit 2 passes; the former thereby screens part of the charge of the nucleus (small circle) for the latter; in turn, the orbit 2 also screens part of the nuclear charge for orbit 1.

the sum of energies of individual electrons, each with its appropriate screened nuclear charge. The degree of screening differs, in general, from one electron to another. Looking at an electron in a given shell and subshell, electrons lodged in higher subshells and shells (i.e., those with higher n or, ones with the same n but higher l) exert little screening effect on that electron, and the screening effect of the rest of the electrons depends on which shells and subshells they are in. It is not always easy to quantitatively work out this screening effect which, moreover, does not constitute a description having a general validity (see sec. 18.5.2.2). In other words, the idea of screening gives us a simplified but incomplete, though useful, description of atomic states and wave functions, and its application differs from case to case.

Since this screening affects the energy levels of the atom as a whole, it has a role to play in determining the frequencies of spectral transitions from various atomic states to others. An example of particular interest relates to *characteristic X-ray spectra of elements*, which I will outline in a section to follow (sec. 18.6).

Problem 18-1

Estimate the ground state energy of a helium atom, on the assumption that each of the two electrons feels a screened nuclear charge of 1.7 times the charge of a proton ($1.6 \times 10^{-19}\text{C}$). Compare your result with the energy that would result if there were no screening.

Answer to Problem 18-1

HINT: The required energy is obtained from the formula (18-4a), which gives the energy of each of the two electrons on putting $Z = 1.7$ (corresponding to a screened nuclear charge of magnitude $1.7 \times 1.6 \times 10^{-19}$ C), and $n = 1$. Multiplying the single-electron energy by 2, the required ground state energy is obtained as $E_g = -2chR \times (1.7)^2$. Using values $R = 1.1 \times 10^7 \text{ m}^{-1}$ (refer to (18-4b), (18-4c)), $c = 3 \times 10^8 \text{ m}\cdot\text{s}^{-1}$, and $h = 6.6 \times 10^{-34} \text{ J}\cdot\text{s}$, we obtain $E_g = -78.7 \text{ eV}$ (approx; $1 \text{ eV} = 1.6 \times 10^{-19} \text{ J}$). The value of the ground state energy obtained without taking into account the nuclear screening is $E_g^{\text{uncorrected}} = -2chR \times (2)^2 = -108.9 \text{ eV}$, which shows that the screening effect is indeed of considerable importance.

NOTE: The value of the helium ground state energy determined by spectroscopic means is $\approx -79 \text{ eV}$.

18.5.2 Quantum theory of atomic states: a brief outline

How does one arrive at the numerical figure expressing the degree of screening of the nuclear charge quoted in problem 18-1? More generally, to what extent can the intuitive ideas on screening sketched in sec. 18.5.1 be justified on more solid grounds provided by quantum theory?

18.5.2.1 The indistinguishability principle and its consequences

Before giving you a brief (and superficial) glimpse into the quantum theory of the atom as a whole, I must say something more on the *indistinguishability* issue. The principle of indistinguishability is a deeply theoretical one, distinguishing quantum theory from classical theory in an essential and fundamental way, and has profound consequences of a practical nature. The consequences all stem from the following corollary to the indistinguishability principle: *for an assembly of indistinguishable particles, there arises definite restrictions on the wave functions describing the possible quantum states of the assembly*, where the restrictions relate to the *symmetry properties* of the wave functions under interchanges of two or more particles belonging to the assembly. In the case of an assembly of electrons (or, more generally, of particles with a *half-integral* value of the spin quantum number), the wave function has to be *antisymmetric* under the interchange of any and every pair of particles in it. The Pauli exclusion principle then

follows as a further corollary from this antisymmetry property.

Closely related to the exclusion principle, the antisymmetry property has the important consequence of causing an *exchange force* to appear between a pair of electrons in an atom described, in an approximate sense, in terms of single-electron states. The wave function describing the joint state of the pair is made up of a *space part* and a *spin part* where the spin part, considered separately, can be either symmetric or antisymmetric depending on whether the *total spin quantum number* of the pair is 1 or 0. The former corresponds to what is referred to as the *triplet* spin state, and the latter to the *singlet* spin state.

Considering a pair of particles with spin quantum numbers s_1 and s_2 , the rules of quantum theory implies that the total spin of the pair is quantized, and the quantum number (S) for the total spin can have values ranging from $|s_1 - s_2|$ to $s_1 + s_2$, successive values of the quantum number differing from each other by unity. In the case of a pair of electrons, for which $s_1 = s_2 = \frac{1}{2}$, the possible values of the total spin quantum number are $S = 0$ and $S = 1$. For any specified value of S , the quantum number M_S describing the component of the total spin angular momentum on any specified axis ranges from $-S$ to $+S$. Thus, for $S = 0$ the only possible value of M_S is zero, while for $S = 1$, one can have $M_S = -1, 0, 1$. This explains the terms 'singlet' and 'triplet', as applied to spin states of the electron pair.

The difference in the symmetry properties of the singlet and triplet spin states results in a difference in the symmetry properties of the corresponding *space parts* of the wave functions of the pair of electrons under consideration, and a consequent difference in the effective charge distributions associated with these space parts. This, in turn, leads to a difference in the electrostatic interaction energies of the pair of electrons in the spin triplet and spin singlet states.

In summary, the indistinguishability principle has the following consequence: the spatial part of the wave function for the spin triplet state ($S = 1$) of a pair of electrons is antisymmetric under an exchange of the electrons, while that for the spin singlet state

($S = 0$) is symmetric, as a result of which the electrostatic interaction energies of the two states differ from one another.

In general, the spin triplet state has a lower energy compared to the singlet state. In the case of two electrons inhabiting the same single-particle state (for which it is necessary that the two have the same values of the quantum number n and also of l), however, the spin triplet state is ruled out by Pauli's exclusion principle.

18.5.2.2 Electron-electron interaction: the central field

Generally speaking, quantum theory gives a reasonably good account of the energies and the stationary state wave functions of the ground states and low-lying excited states of atoms with low and moderate atomic numbers. For this, one needs appropriate approximation schemes where the energies and the wave functions are worked out in several steps of approximation.

The *Hartree theory* is one such scheme with a wide range of applications, where conclusions are obtained on a case-to-case basis, primarily in a numerical form, though a number of conclusions of a general nature are also obtained by comparison of numerical results relating to various different atoms. It is essentially a theory designed to solve the Schrödinger equation for an atom, including the electrostatic interaction energy of the electrons, in a *self-consistent* manner, passing through several *cycles* of computation, where the cycles eventually converge to some desired solution.

To start with, each electron in the atom is assumed to move independently of the others in the field of the nucleus, *and an additional central field* caused by all the remaining electrons, where, in the first cycle, an assumed form for this additional central field is taken by way of an informed guess, satisfying certain general requirements. One then solves the Schrödinger equation numerically so as to obtain a set of energy values and associated single-particle wave functions, and obtains the total energy and the joint wave function of all the electrons by assuming that all the Z electrons in the atom fill up the single-particle energy levels from bottom upwards, in accordance with the exclusion principle. From this, one calculates the effective charge density of the assembly of

electrons (the charge density for a single electron with a wave function $\psi(\mathbf{r})$ is $e|\psi(\mathbf{r})|^2$, where e stands for the electronic charge) and then, a potential function that can be taken to be a *better approximation* than the one assumed at the beginning of the first cycle. A *second* cycle is then carried out, obtaining a still better estimate of the potential (a central, i.e., spherically symmetric one) felt by a single electron in the assembly. Eventually, one achieves a self-consistency where the potential at the beginning of a cycle agrees reasonably well with that coming up at the end of it.

The potential arrived at in this manner includes the Coulomb potential of the nucleus felt by an electron *and* an effective central potential felt by it due to all the other electrons in the atom. It is the latter that can be described as having the effect of partially screening the former where *the degree of screening is found to depend on the principal quantum number n* . The screening description, however, is nothing more than a simplified view of the detailed numerical results of the Hartree theory, and does not enjoy a general validity.

In this simplified picture the screening is found to depend on the occupancy of the single particle levels by the assembly of electrons. In the case of a valence electron (outside the core of filled shells and subshells), its dependence on the quantum number n of the electron can be given a relatively simple interpretation. A simple way to work out the energy of the atom is to keep track of the screening for each electron and then add up the energies of all the electrons in the atom. The idea of screening provides only a partial description of atomic energies and states, since the self-consistent potential resulting from the Hartree theory has a dependence on the distance r of the electron from the nucleus that differs from a Coulomb potential, as a result of which there arises a dependence of the electron energy on the angular momentum quantum number l , which becomes stronger for relatively larger values of l .

The solution to the Schrödinger equation for a potential with a $\frac{1}{r}$ dependence on the distance from the center of attraction has the exceptional feature that the energy of a state depends only on the quantum number n ($\propto \frac{1}{n^2}$ as in the Bohr theory) and not on the quantum number l (there arises, however, a weak dependence on l due to

relativistic effects). For a spherically symmetric potential with a different dependence on r , the energy is determined by both the two quantum numbers.

The dependence of the energy on l in addition to that on n provides the basis of the sequence in which various n, l levels are occupied for ground states of atoms with increasing values of the atomic number Z . Though there does not exist any general formula expressing this dependence for all atoms, the empirically formulated *Madelung rule* works well for a good number of atoms in the first half of the periodic table: subshells and shells are filled up in order of increasing value of $n + l$ while, for equal values of $n + l$, states with lower n are filled up first. For instance, the sequence 1s, 2s, 2p, 3s, 3p, ..., gets modified once the 3p subshell is filled up and, *instead of 3d being filled up next, the two succeeding electrons go to the 4s subshell*.

Coming back to the self-consistent procedure of calculating the central potential, it is by means of a procedure such as the one outlined above that the numerical value of the screened nuclear charge in the case of the helium atom, as quoted in problem 18-1, can be arrived at (the screening is the same for both the electrons in the atom in the ground state of the latter). I repeat, however, that this approach of arriving at the energy of the atom by invoking the concept of screening is nothing more than a convenient simplification, while the more elaborate results of the Hartree theory are to be made use of so as to arrive at results with a greater degree of dependability.

Though the Pauli principle is taken into account in the Hartree theory, the latter does not incorporate the full set of consequences resulting from the indistinguishability principle, which puts a restriction on the symmetry property of the joint wave function describing the state of an assembly of more than one electrons. The theory, however, still remains a useful and convenient one.

18.5.2.3 Electron-electron interaction: the spin-dependent residual term

The self consistent central potential obtained in the Hartree theory constitutes only a first approximation in the description of the atomic structure since it ignores a number

of lesser effects that need to be taken into account for a more complete description. The exchange force arising from the spin-dependent electrostatic interaction between electrons, mentioned in sec. 18.5.2.1, constitutes one such effect where the contribution of the exchange interaction to the energy of the atom arises as a *residual* term in the energy expression. A reasonably good account of the effect of the exchange force is obtained if one considers only the electrons in the outermost orbital(s) of the atom. Considering, for instance, the helium atom, the only occupied orbital in the ground state is 1s, in which the two electrons are in a spin-singlet state (by Pauli's principle; since both the electrons have $n = 0, l = 0$, the two must have antiparallel spins), and there is thus only one possibility for the spatial part of the joint wave function of the electrons (the symmetric one).

Considering, on the other hand, the low lying excited states of the helium atom, one obtains, in the Hartree theory, the first two excited states with configurations 1s2s and 1s2p. As one now considers the exchange interaction in the next order of approximation, the residual term in the interaction energy causes each of these to *split* into two energy levels close to each other, of which the lower one corresponds to the spin triplet ($S = 1$) combination and the upper one to the spin singlet ($S = 0$) combination, as indicated in sec. 18.5.2.1.

18.5.2.4 Spin-orbit coupling: excited states of sodium and magnesium

The self-consistent central field and the spin-dependent residual interaction are but two stages of approximation in the description of atomic states and energies, where other interactions and effects are also of relevance, though to lesser degrees, as a result of which these are to be taken into consideration in subsequent stages of approximation. Even so, no known scheme of approximation is of a general validity over the entire periodic table. In this brief and introductory exposition, I focus on one particular scheme in which the third stage of approximation is provided by what is termed the *spin-orbit coupling*.

Before looking at the spin-orbit coupling effect, here is a useful and convenient approach

of describing the ground states and low lying excited states of a number of atoms that come under the category of 'one-electron' and 'two-electron' ones. In this approach, one refers to the atom of the inert gas appearing immediately before the atom under consideration in the periodic table. An inert gas atom is made up of a set of closed electronic shells around the nucleus whose excited states are separated from the ground state by a considerable gap in the energy scale. Considering the next element in the periodic table, it can be described as a single-electron outside a *core* having the ground state configuration of the inert gas, while the next element in the periodic table can be described as the core together with *two* electrons outside the core. A number of basic principles in atomic structure can be explained in relatively simple terms by referring to atoms of these two categories - the 'one-electron atom' and the 'two-electron atom' referred to above, while the behavior of atoms with more complex structures need much more intricate considerations in quantum theory. For a one-electron or a two-electron atom, the energies of the ground state and low-lying excited states can be obtained as the energy of the core having the structure of an inert gas, plus the low-lying energy levels of the outer electron(s).

The sodium atom is a 'one-electron' atom where there is a core with the configuration $1s^2 2s^2 2p^6$ (the configuration of neon) along with a single electron in the 3s orbital. For this single electron, the residual interaction is not relevant. Moreover, the spin-orbit coupling effect, introduced below, is also not of relevance since it shows up only for electrons with $l \neq 0$ (or for non-zero values of the *total angular momentum quantum number* L and *total spin quantum number* S - see below). In other words, the calculation of the self-consistent central field suffices for a reasonably good description of the ground state of sodium, with the outermost electron lodged in the 3s orbital.

Considering next the first excited state of sodium, it results from the single electron lodged in the 3p orbital outside the core. Spectroscopic observations reveal that this corresponds to not one but *two* energy levels lying close to each other - a phenomenon referred to as *doublet splitting* (the term is more commonly used to indicate the occurrence of a pair of neighboring lines in the emission or absorption spectrum of sodium - see below).

This is due to the fact that the electron, in the course of its orbital motion around the nucleus, effectively feels a magnetic field (the electric field of the nucleus observed in the rest frame of the latter, when viewed from the frame of reference of the revolving electron, is transformed to an admixture of an electric *and* a magnetic field resulting from the Lorentz transformation of the frame of reference). The interaction of the magnetic dipole moment of the electron, arising due to its spin, with this magnetic field, referred to as the *spin-orbit coupling*, is responsible for the two energy levels mentioned above.

The stationary states of the atom resulting from the doublet splitting of the 3p state of the single electron outside the core differ in the *total angular momentum* composed of the orbital angular momentum (described by quantum number $l = 1$) and the spin angular momentum (quantum number $s = \frac{1}{2}$). The principles of quantum theory demand that this angular momentum be also quantized like the spin and the orbital angular momenta, corresponding to which one has the quantum number j that can have values ranging from $|l - s|$ to $l + s$, successive values differing from each other by unity. In the case of the 3p state of the outermost electron of sodium, the possible values of j decreed by this rule are $j = \frac{1}{2}$ and $j = \frac{3}{2}$, corresponding to the two neighboring levels mentioned above, produced by the spin-orbit coupling.

Transitions from these two levels to the 3s ground level result in the so-called D-lines - two lines in the emission spectrum (as also in the absorption spectrum) of sodium whose frequencies are close together.

The element occurring next to sodium in the periodic table is magnesium, which is a 'two-electron' atom having, in its ground state two 3s electrons outside the core, for which the exchange interaction is not of relevance, analogous to the case of the helium atom. The spin-orbit coupling is also not of relevance for the ground state of magnesium because of the fact that the quantum numbers corresponding to the *total spin* and *total orbital angular momentum* in the ground state are both zero.

Coupling of angular momenta.

It is important to understand how the various angular momenta of the electrons in an atom are to be combined in a *relevant* way. Each of the electrons has its own spin and orbital angular momenta, described by quantum numbers s, l . For a ‘two-electron’ atom such as magnesium one has four angular momenta to consider, described by quantum numbers s_1, s_2, l_1, l_2 (in this, the angular momenta of the electrons comprising the core need not be considered, since these add up to zero value). Of the possible compositions of various combinations of two or more of these angular momenta, only those are relevant which remain *conserved* (or, at least, approximately conserved) as dynamical variables, because it is the set of these angular momenta that are relevant in describing the stationary states (the ground state and the excited states) of the atom. For instance, in describing the stationary states resulting from the spin-orbit coupling outlined above, the angular momenta of relevance are the resultant spin angular momentum and the resultant orbital angular momentum, for which the quantum numbers are denoted by S and L (upper case letters; lower case letters are used to denote the quantum numbers for individual electrons). The principle of composition of angular momenta in quantum theory tells us that, if q_1, q_2 be the quantum numbers for any two angular momenta, then the quantum number q corresponding to the composition of these two can have values ranging from $|q_1 - q_2|$ to $q_1 + q_2$, with successive values differing by unity.

The low-lying excited states of magnesium result from one of the two outer electrons lodged in the 3s state and the other in the 3p state, for which one has $L = 1$ (the only possibility corresponding to $l_1 = 0, l_2 = 1$), while two values are possible for the total spin quantum number, namely $S = 0, 1$ (reason out why; here $s_1 = s_2 = \frac{1}{2}$). Of these the former is a singlet state (the projection of the total spin angular momentum on any specified direction can have only one possible value), while the latter is a triplet (the projection can have three possible values). According to what has been indicated in sec. 18.5.2.1, the triplet state has a lower energy as compared with the singlet, in virtue of the exchange interaction between the two electrons.

The spin-orbit coupling, which one has to consider in the next level of approximation,

comes into play for the triplet state (with $L = 1, S = 1$), where it results in *three* levels, with energies close together, differing in their *total angular momentum quantum number* J (resulting from the composition of the total spin and the total orbital angular momentum) whose possible values are 0, 1, 2 (reason out why). These (in order of increasing energy) constitute the first three excited states of magnesium. In the next excited state, namely the singlet state with $L = 1, S = 0$, the spin-orbit coupling is again of no relevance since, for this state, the total angular momentum quantum number can have only one possible value ($J = 1$).

1. In the above considerations relating to the ground states and low-lying excited states of atoms, I have not made reference to the spin and orbital angular momenta of the electrons belonging to the core since, as mentioned above, the quantum numbers for the total spin and orbital angular momenta of a core made up of closed shells and subshells turn out to be zero, where the core is assumed to be in its ground state.
2. The above scheme of approximation (referred to as the L - S coupling scheme), where the spin-orbit coupling effect is small compared to the spin-dependent electron-electron interaction, is not the only possible one. For some atoms with relatively large atomic numbers, the spin-orbit coupling has a larger effect on the state of the atom, as compared with that of the spin-dependent residual interaction. A different scheme of approximation, termed the j - j coupling scheme, gives a better result for these. In many instances, however, none of the two approximation schemes works, and more complex considerations are needed to describe their ground states and the low-lying excited states.

Problem 18-2

Referring to Bohr's theory, estimate the size of a sodium atom.

Answer to Problem 18-2

The radius of a sodium atom can be estimated by referring to the Bohr formula (eq. 18-5). In this formula, we take $n = 3$, the principal quantum number of the valence electron, and $Z_{\text{eff}} = 1$, the screened nuclear charge that determines the effective field felt by the valence electron due

to the screening of the nuclear charge by the core electrons. Substituting known values of ϵ_0 ($\approx 8.85 \times 10^{-12}$), h ($\approx 6.62 \times 10^{-34}$), m ($\approx 9.1 \times 10^{-31}$), e ($\approx 1.6 \times 10^{-19}$) (all in SI units), one obtains $r \approx 4.77 \times 10^{-10}$ m (approx).

NOTE: This is an overestimate since it does not refer to the maximum density of the effective charge distribution around the nucleus, and the assumed value of the screened nuclear charge is incorrect. A more realistic calculation based on quantum principles gives $r \approx 1.9 \times 10^{-10}$ m. However, the Bohr formula gives a much better estimate for the *relative* radii of single-electron atoms, such as sodium and potassium.

18.5.3 The atom as a whole: summary and overview

The atom is a complex object, where the complexity arises due to the fact that, generally speaking, it contains *several interacting constituents*, i.e., belongs to the class of what are referred to as *many-body* systems. There does not exist any general quantitative theory where the stationary states of an arbitrarily chosen atom, along with their energies, can be determined accurately without introducing appropriate simplifying assumptions and approximations.

A common starting point in any such scheme of approximation is the calculation of an effective central (i.e., spherically symmetric) potential in which the electrons can be assumed to move independently of the others (subject to the restrictions of the Pauli principle) where the potential includes the Coulomb field of the nucleus *and* the that resulting from the Coulomb interaction among the electrons. One widely applied procedure for this is the self-consistent calculation in the Hartree theory.

A considerable simplification in the description of the ground states and the low-lying excited states of atoms results if one looks at the atom as a core with an inert gas structure, together with a number of outer electrons (referred to as the valence electrons), where now the possible joint states of these outer electrons assume relevance in the explanation of numerous properties of the atom. Here a number of additional factors, to be taken into account in successive stages of approximation, are the spin-dependent

electron-electron interaction and the spin-orbit coupling, between the outer electrons. In the case of relatively simple lighter atoms, the former commonly dominates over the latter, and the $L - S$ coupling scheme is found to explain the ground states and the low-lying excited states of the atom. For certain heavy atoms, on the other hand, another scheme of taking into account the above-mentioned effects is the so-called $j-j$ coupling scheme. Both these schemes however, involve simplifying assumptions, and one needs, more generally, methods of calculation of a more complex nature, where analytical methods are to be combined with a numerical approach of computation so as to yield meaningful results.

18.6 Continuous and characteristic X-ray spectra

If a stream of electrons possessing sufficient kinetic energy (by means of an accelerating voltage applied to them) is made to hit a metal target or one made up of a heavy element in an evacuated tube, something pretty interesting is found to happen: electromagnetic waves of short wavelength, commonly known as X-rays, are emitted from the target. One mechanism responsible for this phenomenon is the deceleration of the electrons, which give up their energy to the atoms making up the target material and, in the process, emit part of the energy in the form of electromagnetic waves. The mechanism at work is similar to the one operative in a broadcasting antenna where the alternate acceleration and deceleration of charges in a wire or a rod carrying an alternating current results in the emission of radio waves - electromagnetic waves of longer wavelength.

These X-rays are found to be made up of components with frequencies distributed continuously over a certain range, i.e., in other words, are characterized by a *continuous spectrum*, and do not depend much on the target material being used. However, there is *another* mechanism at work, by which X-rays of *specific* frequencies are emitted, these being characteristic of the target material. These components are said to constitute the *characteristic spectrum* of the emitted X-rays. They originate in *transitions between atomic states* of the element the target material is composed of.

The characteristic spectrum is made up of *lines*, i.e., specific frequencies, that can

be grouped into a number of series, similar to the line spectrum of hydrogen atom. Spectroscopists have their own symbols for designating these lines, like K_α , K_β, \dots , L_α , L_β, \dots , and so on.

Fig. 18-4 depicts schematically the X-ray spectrum obtained by analyzing the emission from a X-ray tube with a target made of a pure metal where the continuous spectrum is shown along with a pair of sharp lines belonging to the characteristic spectrum. The plot in this figure is of the X-ray energy radiated per second per unit wavelength ($I(\lambda) = \frac{\delta P}{\delta \lambda}$, where δP is the power radiated within a small wavelength interval $\delta \lambda$ around the wavelength λ) as a function of the wavelength (λ). Note from the figure that the continuous spectrum is terminated on the short wavelength side at a certain wavelength λ_{\min} , where λ_{\min} depends on the maximum energy (E_{\max}) of the stream of electrons used in the X-ray tube to hit the target electrode (in practice, the energies of the electrons are distributed over a narrow range and may, for the sake of simplicity, be assumed to be the same for all the electrons; this energy E depends on the voltage through which the electrons are accelerated before they are made to hit the target).

The minimum wavelength corresponds to the maximum energy of the photons emitted from the target, which in turn equals the maximum energy of the electrons hitting the target, where the entire energy (E) of an electron gets converted into the energy of an X-ray photon. On the other hand, if only a part of the electron energy E gets converted to the photon energy (the rest being dissipated as heat in the target), then the photon will be characterized by a wavelength larger than λ_{\min} .

The relation between the energy (E) of a photon and the wavelength (λ) is obtained from the equations (16-42) by making use of the relation $\lambda = \frac{c}{\nu}$ between the frequency and the wavelength. One thereby obtains

$$\lambda_{\min} = \frac{hc}{E_{\max}}. \quad (18-6)$$

Referring to the lines making up the characteristic spectrum, Mosley made a systematic

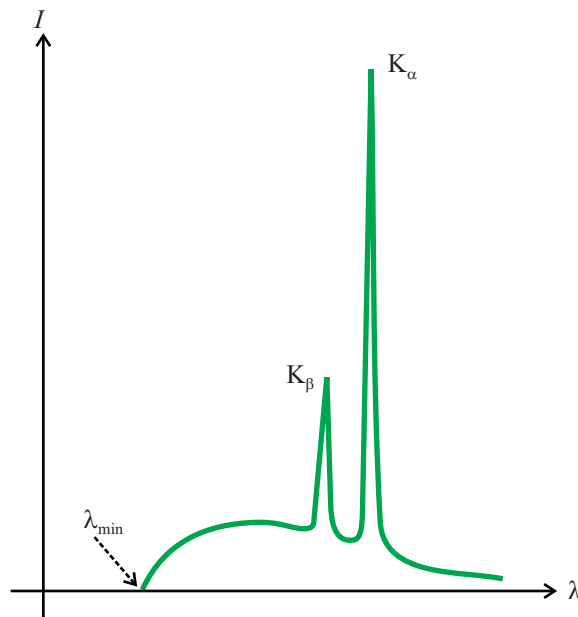


Figure 18-4: Variation of power per per unit wavelength with wavelength λ (schematic) for X-rays emitted from an X-ray tube; the continuous spectrum along with two sharp lines belonging the characteristic spectrum is shown; the continuous spectrum is terminated on the low wavelength side at a certain wavelength λ_{\min} .

analysis of the frequencies of these lines, and came up with a neat formula. He demonstrated that the frequencies of the K_α line, for instance, depends in a simple manner on the *atomic number* of the element making up the target material.

18.6.1 Bohr's theory and X-ray spectra

Consider now an atom with atomic number Z with its electrons in the various shells around the nucleus and assume that an electron has, by some means, been knocked out of the K shell, i.e., the one nearest to the nucleus, corresponding to the principal quantum number $n = 1$. This requires a supply of energy from outside like, for instance, the energy imparted by another energetic electron speeding past the atom (recall how a material can be made to emit X-rays by getting a beam of energetic electrons to hit on it). Since the K shell contains two electrons, a deficit of one electron in it means that it now contains just one electron. As a consequence, it is now possible for one other electron to make a transition from an outer shell to this depleted K shell.

For instance, imagine an electron from the L shell to make a transition to the K shell. This is similar to the transition of the electron from one stationary state to another in a hydrogen atom where a photon is emitted from the atom and one may, just for the fun of it, try to work out the frequency of the photon by making use of the Bohr formulae (18-4a) and (16-40b). However, before we do this, we have to recall the phenomenon of *screening* of the nuclear charge mentioned in section 18.5.1. From the point of view of the electron making the transition from the L to the K shell, the single electron in the partially filled K shell does the screening and since the latter is lodged in an interior shell, the reduction in the nuclear charge due to the screening is likely to be by the full magnitude of the electronic charge.

In other words, the electron in the L shell effectively feels a nuclear charge $(Z - 1)e$ and, in the Coulomb field of this screened charge, makes a transition to the K shell. Its initial and final energies are then given by the expression (18-4a) with $n = 2$ and $n = 1$ respectively, *and with Z replaced with $(Z - 1)$* . The frequency of the photon emitted in such a transition is obtained by dividing the energy difference by the Planck constant:

$$\nu = cR(Z - 1)^2 \left(\frac{1}{1^2} - \frac{1}{2^2} \right) = \frac{3}{4}cR(Z - 1)^2. \quad (18-7)$$

This is found to agree with a formula, relating to the K_α frequencies of various elements, written out by Mosley on the basis of his analysis of X-ray data. Mosley's findings were of a more general nature and included predictions on the frequencies of a number of other lines in the characteristic spectrum as well. For instance, a K_β line corresponds to a transition from the M shell ($n = 3$) to the K shell with a vacancy in the latter. In accordance with the above approach of making use of Bohr's theory along with the basic idea on screening of the nuclear charge, the frequency for such a transition is given by the formula

$$\nu = \frac{8}{9}cR(Z - 1)^2, \quad (18-8)$$

where, once again, the equations (18-4a) and (16-40b) have been made use of, with $Z - 1$ replacing Z and, and with the initial and final quantum numbers taken to be 3 and 1

respectively. Formulas of the type (18-7) and (18-8) taken together are said to constitute *Mosley's law* since these are found to be in agreement with Mosley's observations on X-ray spectra.

The law is commonly expressed in the form of a linear relationship between the square root of the frequency of characteristic X-rays belonging to any given series, and the atomic number Z of the element from which the characteristic X-ray is emitted, i.e., one of the form

$$\sqrt{\nu} = a(Z - b), \quad (18-9)$$

where a and b are constants, with the former being a characteristic of the series under consideration and the latter that of the metal from which the x-ray is emitted, being determined by the degree of screening of the nuclear charge.

Continuing with examples of characteristic X-ray lines, an L_α line arises by virtue of a transition from an M shell to an L shell, with a vacancy in the latter caused by an electron having been ejected from it by means of a collision. All these characteristic lines can be arranged into a number of series analogous to the lines making up the optical spectra of lighter atoms. The comparatively large atomic numbers of metallic targets results in the frequencies, given by formulas of the type (18-7) and (18-8), falling in the X-ray region of the electromagnetic spectrum.

Fig. 18-5 depicts schematically the transitions responsible for the K_α and K_β lines in the X-ray spectrum of an element, where the energy levels corresponding to the various shells are represented by horizontal lines, with the energy increasing from the lower to the higher levels.

Problem 18-3

Electrons accelerated by a potential difference of 2.0 kV are made to hit a target in an X-ray tube. Find the minimum wavelength characterizing the X-rays emitted from the target.

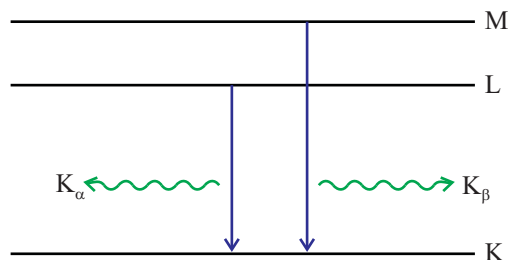


Figure 18-5: Depicting the transitions responsible for the K_α and K_β lines in the X-ray spectrum of an element; the K_α line arises due to the transition from a state in the L-shell to one in the K-shell, while the transition responsible for a K_β line occurs from the M-shell to the K-shell; in either case, a vacancy is required to be formed in the K-shell; the remaining electron in the K-shell causes a screening of the nuclear charge felt by the electron making the transition; in the diagram, the energy increases from the bottom upward.

Answer to Problem 18-3

HINT: A potential difference of 2.0 kV corresponds to an energy 20 keV, i.e., $E = 2.0 \times 10^3 \times 1.6 \times 10^{-19}$ J. The minimum wavelength λ_{\min} (corresponding to a frequency $\nu_{\max} = \frac{c}{\lambda_{\min}}$) is obtained by equating this to the photon energy $h\nu_{\max}$. Thus, $\lambda_{\min} = \frac{hc}{E}$, which works out to 0.62 nm.

Problem 18-4

The K-, L-, and M shell energies for tungsten ($Z=74$) are, respectively, -69.5 keV, -11.3 keV, and -2.3 keV, where it is assumed that there is a vacancy in the K shell of the atom. What are the wavelengths of the K_α and the K_β lines emitted from tungsten. Estimate, from these, the wavelengths of the corresponding lines of Molybdenum ($Z=42$).

Answer to Problem 18-4

HINT: The wavelength corresponding to a transition from a level with energy E_1 to one with energy E_2 is $\lambda = \frac{hc}{E_1 - E_2}$. Thus, the wavelength λ_α of the K_α line of molybdenum will be $\lambda_\alpha = \frac{6.63 \times 10^{-34} \times 3 \times 10^8}{(69.5 - 11.3) \times 10^3 \times 1.6 \times 10^{-19}}$ m = 0.021 nm (approx). According to Mosley's law, the wavelength of the K_α line of molybdenum is related to that for tungsten as $\frac{\lambda_\alpha'}{\lambda_\alpha} = \frac{(74-1)^2}{(42-1)^2}$ (reason this out), which gives the estimate $\lambda_\alpha' = 0.068$ nm. In a similar manner, the K_β wavelength for tungsten is obtained by taking $E_1 - E_2 = (69.5 - 2.3)$ keV, which gives $\lambda_\beta = 0.0185$ nm, and the corresponding wavelength for molybdenum will be $\lambda_\beta' = 0.059$ nm.

18.7 Atomic spectra

With a knowledge of the stationary states of atoms, one can seek an explanation of *atomic spectra*. The first successful explanation of an atomic spectrum was that by Bohr, based on early quantum ideas, where the latter were mixed with classical concepts as well, though in a clever manner so as not to stand in the way of arriving at meaningful results. Later developments in the quantum theory of the atom paved the way for a more complete understanding of atomic spectra in terms of the stationary states and energy levels of the atom, the basic ideas relating to which have been briefly introduced in sec. 18.5.

An atom can continue in a stationary state of energy, say E only so long as it is not disturbed by an external influence. However, external influences are inevitable - specifically, from the electromagnetic field around it. Such an electromagnetic field is produced not only by sources outside the atom, but in the form of the *vacuum field* as well.

Considering the atom together with the electromagnetic field, the ground state of the composite system consists of the atom in its ground state, and the electromagnetic field in its *vacuum state*, where the latter is the one in which the field is depleted of photons. While such a state of the composite system is a stable one, any *other* state is potentially unstable. For instance, if the atom be in an excited state of energy E , its interaction with the electromagnetic field may cause it to make a *transition* to a lower energy state of energy E' , emitting a photon (which appears as an excitation of the electromagnetic field) of energy

$$\nu = \frac{E - E'}{h}, \quad (18-10)$$

(recall, in this context, fig. 16-7, and eq. (16-40b)). This is how *spontaneous* emission takes place from an atom, where the term spontaneous signifies that the emission process occurs without the participation of photons in the electromagnetic field surrounding the atom. In other words, such an emission process occurs with the electromagnetic field initially in its vacuum state ending up finally in a state with an emitted photon.

In addition, there may occur *stimulated* emission as well where the presence of one or more photons in the electromagnetic field modifies the probability of the atom making a transition from a higher to a lower energy state (recall the introduction to spontaneous and stimulated emission in section 15.6 of which the present discussion is, in the main, a repetition). Finally, the interaction between the atom and the electromagnetic field can give rise to *absorption* of a photon, with the atom making a transition from a lower to a higher energy state.

The emission and absorption processes give rise to the emission and absorption *spectra* of the atom, either of which consists of a characteristic series of frequencies (refer to eq. 18-10). When one considers the emission or absorption spectrum of a material made up of atoms in aggregate, there may arise a continuous distribution of frequencies as well.

With a knowledge of the stationary states of the atom and of the way the atom interacts with the electromagnetic field, one can invoke the rules of quantum theory so as to work out the *transition probabilities* between various atomic states involved in the emission and absorption processes. On going through such exercises one ends up with the following pertinent information: (i) the set of characteristic frequencies making up the emission and absorption spectra of an atom, (ii) the relative *intensities* of the various characteristic spectral lines, (iii) the states of polarization of the photons radiated in an emission process (see, in this context, sec. 15.6.2), and finally, (iv) *selection rules*, if any, characterizing the transitions.

Of these, the selection rules may be looked upon as special instances of the intensity rules resulting from the transition probabilities that can be worked out from the basic quantum principles. Depending on the *quantum numbers* characterizing the initial and final states in a transition, it may be found that the transition probability works out to a zero or a very low value. The frequency corresponding to such a transition will then be absent from the spectrum of the atom under consideration. The existence of selection rules happens to be a consequence of certain *symmetry* properties of the interaction between the atom and the electromagnetic field, and associated *conservation* principles.

However, I will not enter here into a more detailed explanation of these principles.

18.8 Physics of the atomic nucleus

18.8.1 The atomic number and the mass number

The atomic nucleus is a tiny and dense structure forming the core of the atom. While almost the entire mass of the atom is concentrated in the nucleus, the volume occupied by the latter is negligible compared to that of the atom.

The nucleus is made up of neutrons and protons, where a common name for these two is a *nucleon*. The use of a common name is suggestive of an identity, at a deeper level, between these two particles. This I will come to later, in sections 18.8.2 and 18.8.9.

As already mentioned, the number of nucleons (A) making up a nucleus is termed its mass number since it determines, to a good degree of approximation, the mass of the nucleus (as also of the entire atom) in accordance with the formula (18-2b). The number of protons (Z), on the other hand, is termed the atomic number. It determines the nuclear charge as also the total charge of the extra-nuclear electrons in the neutral atom. More importantly, it is the atomic number that determines the structure and properties of the atom as whole, as outlined in earlier sections. The number of protons in relation to the number of neutrons in the atomic nucleus is also relevant in determining the *binding energy* of the latter or, put differently, the *stability properties* of the nucleus (see sec. 18.8.4).

18.8.2 The nucleon: internal characteristics

In section 18.3.3 the spin was introduced as an *internal* characteristic of an electron. Other particles like the proton and the neutron are also characterized by their spins where the spin signifies an angular momentum of an intrinsic nature. Depending on the magnitude of this angular momentum, the angular momentum vector can have only certain discrete orientations in space, analogous to the space quantization of the atomic orbitals. The spin angular momentum of a proton or a neutron resembles that of an

electron in that the spin can have just two orientations in space where one refers to the corresponding quantum states as the ‘spin up’ and ‘spin down’ states respectively.

On the face of it, a spin-up electron and a spin-down electron could be looked upon as two different particles. However, the two are found to behave in an entirely identical manner in most physical situations, differing only in a few respects such as their response to magnetic fields. This is why the two are identified as two states of the same basic entity, namely, the electron. The same applies to the two spin states of the proton or of the neutron.

However, the proton and the neutron themselves behave in a notably similar manner in a number of crucial respects, so much so that their difference can, in a sense, be likened to the difference between a spin-up and a spin-down electron. In other words, one can look upon a proton and a neutron as two different *states* of the same basic entity, a single particle if you like, the *nucleon*. Analogous to the fact that the spin-up and the spin-down states of an electron differ in respect of the orientation in space of the spin angular momentum, the proton and the neutron states of a nucleon can be described as differing in the orientation, in an *internal space*, relating to a similar intrinsic property referred to as the *isospin*. Thus, the proton can be described as an isospin-up nucleon, while the neutron is an isospin-down nucleon.

At a more fundamental level, the distinction between a proton and a neutron is explained in terms of *quarks*, the building blocks of elementary particles we will have a look at in section 18.8.9.4.

The nucleons are bound together in the nucleus by binding forces referred to as *nuclear* forces, which belong to a broader class of forces termed the *strong interaction* (see sec. 18.8.3). It is characterized by the crucial feature that protons and neutrons behave identically so far as their interactions by means of these strong forces are concerned. In other words, protons and neutrons cannot be distinguished by referring to phenomena based on the strong interaction forces alone. They do, however, differ when phenomena involving other types of interactions, like the electromagnetic ones, are considered.

18.8.3 The interaction force between nucleons

The force of interaction between nucleons, responsible for binding the nucleons in a nucleus, have a number of characteristic features not found in forces of other types, such as those of electromagnetic origin.

One of these features I have already mentioned above - strong interactions do not distinguish between protons and neutrons. The interactions, moreover, are of a *short range*, the typical distance over which the strong interaction forces operate being of the order of a *fermi* ($fm, \equiv 10^{-15} \text{ m}$). Since the nucleons are bound to one another within this range, the linear dimension of the nucleus is also of this order (see equations (18-11), (18-12) below). By contrast, electrons are bound to a nucleus by electrical forces which are of a much longer range. The typical dimension characterising atomic orbitals is $\sim 10^{-10} \text{ m}$, which means that the size of the nucleus is $\sim 10^{-15}$ times the over-all size of an atom. Since the mass of the atom is, to a good degree of approximation, the same as that of the nucleus, the density of nuclear matter is $\sim 10^{15}$ times the typical density of macroscopic solid materials, for which the ratio of mass and volume is typically of the order of the corresponding ratio for an atom.

The fact that the nuclear force is a short range one means that while the force between two nucleons is strong within a distance of $\sim 1 \text{ fm}$, it falls off very rapidly as the distance is made to increase. Fig. 18-6 depicts schematically the potential energy diagram characterizing the nuclear force, where the variation of the potential energy (U) of a nucleon pair is plotted against their separation (r) (recall that the force is proportional to the rate of change of potential energy, taken with a negative sign; an increasing function $U(r)$ represents an attractive force). One infers from this figure that the force decreases sharply over a distance R (typically, of the order of R_0 in eq. (18-12)). Incidentally, the figure depicts the variation of what determines the *central* component of the nuclear force, while the latter happens to have a *non-central* component as well. The graph in fig. 18-6 looks like a trough or a well, and is at times referred to as a 'potential energy well', the 'depth' of the well shown in the figure being U_0 .

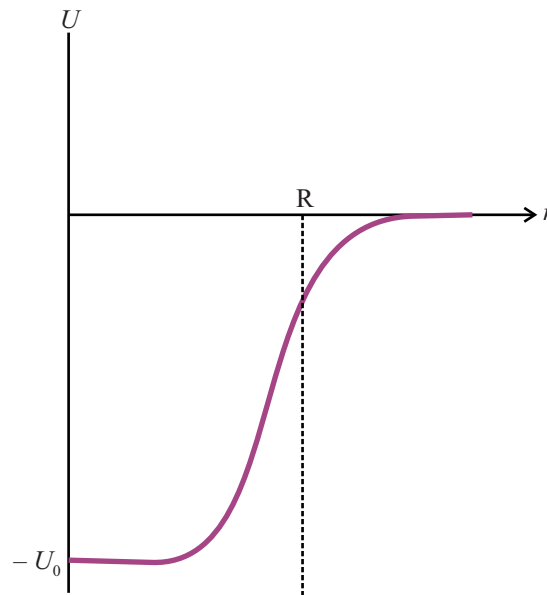


Figure 18-6: Nuclear force potential well of depth U_0 ; the potential energy $U(r)$ of a nucleon pair is plotted against their separation r ; the edge of the well rises sharply at a distances of the order of R , the range of the potential; a similar graph depicts the variation of the potential energy of a nucleon in a nucleus, in which case R stands for the nuclear radius; with reference to the inter-nucleon potential energy, however, R is of the order of R_0 in eq. (18-12).

The nuclear forces are, moreover, *strong* in the sense that these are several orders of magnitude stronger than the electromagnetic forces, the nuclear force between a pair of protons at a distance ~ 1 fm from each other being $\sim 10^2$ times larger than the electrical force between them.

In the description and analysis of interactions between microscopic particles, the concept of *interaction energy* is more appropriate than that of *force*. Instead of talking of the force between two nucleons separated by a distance of ~ 1 fm, one thus talks of the *binding energy* of a system made of two nucleons. In this context, recall the introduction to bound systems and binding energy in section 16.13.3. Incidentally, the above comparison does not refer to the forces that confine the *quarks* (see sec. 18.8.9.4 below) in a nucleon, which is a more basic indicator of the force of *strong interaction* (sec. 18.8.9.5).

One other characteristic of nuclear forces is that these are *spin-dependent* and, as a

consequence, have a *non-central* component. Considering a pair of nucleons at a separation within the characteristic range of the strong interactions indicated above, their interaction energy depends on the *relative orientation of their spins*. More specifically, the interaction energy is made up of two parts - one independent of the spin orientations and the other depending on the relative spin orientation of the nucleons. Accordingly, the force between the nucleons may be said to consist of two parts - a central, and a non-central one (see section 3.17.2 for an introduction to these concepts). While the line of action of the central component lies along the line joining the two nucleons, as in the case of the electrical force between two protons, the line of action of the non-central component depends on the spin orientations of the nucleons.

Finally, nuclear forces exhibit what is referred to as the *pairing interaction*. A pair of protons with oppositely directed spins (i.e., one with its spin up, and the other with spin down) forms a strongly bound structure as compared to a pair with the spins pointing in the same direction. The pairing interaction, which results from the spin dependence of the nuclear force, is a quantum mechanical feature and explains a number of important nuclear phenomena.

18.8.3.1 The saturation property of nuclear forces

A nucleus is made up of a number of nucleons bound to one another by strong forces. Considering any one of these nucleons, it can feel the strong force exerted by other nucleons in its immediate neighborhood. The effect of all the *other* nucleons, farther away compared to the neighboring ones, can be ignored by virtue of the fact that the nuclear forces fall off rapidly with distance (recall figure 18-6). Fig. 18-7 shows schematically a number of nucleons in a nucleus, with the nucleons in the immediate neighborhood of the nucleon marked A being identified with the help of dotted lines connecting the latter to these neighboring nucleons. These dotted lines indicate that the nucleon A interacts with these neighboring nucleons by the strong interaction forces. One can describe this by saying that *bonds* have been established between A and its neighboring nucleons, while no such bonds are formed between A and the other nucleons farther away from it. In a manner of speaking, these bonds are similar to the ones binding the atoms in a

molecule in that the atoms are bound together in a molecule just like the nucleons in a nucleus, though the forces between atoms are not always short-range ones.

If the nucleon (A) under consideration is surrounded on all sides by neighboring nucleons, i.e., it is bound to the maximum possible number of neighboring nucleons interacting with it from all sides, then one can say that all the possible bonds that the nucleon can form are used up, or *saturated*. On the other hand, if a nucleon is not surrounded equally on all sides by other nucleons, then all its bonds are not saturated, and it is bound to a lesser extent in the nucleus compared to the one with all its bonds saturated.

Unless a nucleus is a light one, with relatively few nucleons in it, most of the nucleons in the nucleus are fully saturated like the nucleon marked A in fig. 18-7, and only a few nucleons like the one marked B in the figure have a different kind of surroundings, with some of their bonds left unsaturated.

Ignoring these relatively few nucleons with unsaturated bonds, a nucleus can then be looked upon as an assembly of nucleons all similarly bound to one another, with all the bonds of each of the nucleons saturated by other, neighboring ones. Since all the nucleons behave identically and therefore must have, on the average, identical surroundings, being equally populated on all sides with similar other nucleons, the nucleus can be assumed to be a *homogeneous* distribution of nucleons.

This would have been a good description of the nucleus, if the relatively few nucleons like the one marked B in fig. 18-7 were not there. In reality, however, these nucleons with some of their bonds unsaturated, do exist, being the ones near the *surface* of the nucleus. The inter-nucleon forces being short range ones, one can imagine a thin surface layer of nucleons which are of this type. The remaining nucleons, constituting a majority are, by contrast, of the saturated type.

This, then, constitutes a simple description, which is a useful and convenient one unless the nucleus belongs to the group of light nuclei in which the surface nucleons,

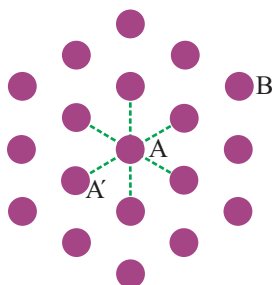


Figure 18-7: Saturation property of nuclear forces; an assembly of nucleons is shown where the assembly takes up the shape of a sphere; the nucleus marked A has all its 'bonds' saturated, i.e., interacts with a maximum number of other nucleons filling up its immediate neighborhood, its interaction with nucleons farther away being negligible; the situation is essentially the same for any other nucleon, like A' in the interior of the assembly; the nucleon B has neighbors only on one side of it and can be described as one whose bonds are not saturated; for large nuclei, the effect of these unsaturated nucleons on the nuclear binding is negligible.

with unsaturated bonds are, in relative terms, no less numerous than the ones with saturated bonds.

18.8.3.2 The nucleus as a liquid drop: nuclear radius

Looked at this way, a nucleus resembles a *liquid drop* because the intermolecular forces in a liquid drop are also characterized by a similar saturation property. The molecules in the liquid drop are all neutral ones and the effective forces between them are of a short range compared to the relatively longer range electrical forces between charged particles. For instance, the fluctuations in the charge density in one molecule may momentarily induce a dipole moment in a neighboring one, and the intermolecular forces may arise due the interaction between these transitory dipole moments. A mathematical analysis of such forces show that, in spite of being of an electromagnetic origin, these are indeed of an effectively shorter range compared to the electrical force between charged particles.

As a result of such short range forces between the molecules of a liquid drop the molecules, in course of their motion, attain a configuration where the drop becomes a *spherical* one unless its shape is disturbed by external influences. In such a configuration, most of the molecules have all their bonds saturated, i.e., are equally surrounded on all sides by neighboring molecules and the molecules are, on the average, distributed homogeneously in the drop. The molecules near the surface of the drop, however, have

their bonds unsaturated and are less bound compared to the molecules in the interior of the drop. Indeed, the entire assembly of molecules tends to assume the most bound configuration, which is why the drop tends to become spherical in shape because the spherical shape minimizes the number of surface molecules compared to those in the interior of the drop.

If we now consider two drops made of the same molecules and assume that the effect of the surface molecules can be ignored, then the molecules in each drop will have identical surroundings since each of these are equally surrounded by other molecules, and thus the mass distributions in the two drops will be identical. In other words, the density of a liquid drop will be independent of its size, which is indeed what is observed in reality.

An analogous conclusion holds for nuclei as well. Since all nuclei are made of identical nucleons and the inter-nucleon interactions are characterized by the saturation property, the density of nuclear matter has to be the same for all nuclei to a good degree of approximation. In other words, the volume of a nucleus has to be proportional to its mass, i.e., to its mass number A (see eq. (18-2b)). In addition the nucleus, like the liquid drop, has to be of a spherical shape whose volume is proportional to R^3 , where R stands for the nuclear radius. One then arrives at the conclusion that the nuclear radius R is related to its mass number A as

$$R^3 \propto A, \text{ or } R = R_0 A^{\frac{1}{3}}, \quad (18-11)$$

where R_0 is a constant whose value is the same for all nuclei. Experimental observations have led to the value

$$R_0 \approx 1.3 \text{ fm} = 1.3 \times 10^{-15} \text{ m}. \quad (18-12)$$

Thus, for instance, a nucleus with $A = 64$ will have a radius of nearly 5 fm.

Speaking of the spherical shape of a liquid drop and of a nucleus, it is not sufficient to

just point at the saturation property of the forces operative in the two cases. What one needs in addition, is the *mobility* of the constituent particles - the molecules in one case and the nucleons in the other. It is by virtue of this mobility that the constituents can seek out a stable configuration - the spherical one - that minimizes the *free energy* of the system.

18.8.4 Nuclear binding energy and mass: nuclear stability

Consider a number of protons and neutrons bound together to form a nucleus and compare this with another configuration of the same nucleons, namely the one where these are separated from one another by large distances with each of these nucleons being, moreover, at rest. Calling these the bound and the separated configurations respectively, one can compare the *energies* of these two configurations.

18.8.4.1 The mass-energy equivalence principle

In doing this, we will make use of *Einstein's mass-energy equivalence relation* (refer to sec. 17.2.8.6),

$$E = mc^2. \quad (18-13a)$$

This relation follows from the basic principles of the special theory of relativity and involves a re-interpretation of the concepts of mass and energy which is found to be necessary for the consistency of the theory. For a particle whose velocity (v) is small compared to the velocity of light in vacuum (c ; such a particle is referred to as a *non-relativistic* one), it reduces to

$$E = m_0c^2 + \frac{1}{2}m_0v^2, \quad (18-13b)$$

where m_0 is the mass of the particle as observed in a frame of reference in which it is at rest, and is termed as its *rest mass*. This relation is consistent in that it expresses the energy of a non-relativistic particle as the sum of its rest energy and its kinetic energy. The latter is given by the familiar expression $\frac{1}{2}m_0v^2$ where one notes that what one

commonly refers to as the mass of the particle, is nothing but its rest mass. Incidentally, with the mass-energy equivalence principle in mind, one often refers to the energy (E) of a system by just specifying its mass (m), expressing the energy in mass units (kg), the corresponding value in energy units (J) being obtained by multiplying with the constant c^2 . Conversely, the mass of an unbound or a bound (i.e., composite) particle is often expressed in energy units.

18.8.4.2 Units for nuclear masses

While the kg is the basic SI unit of mass, the masses of microscopic particles like the nuclear particles (proton and neutron) and of nuclei are commonly expressed in the more convenient *atomic mass unit* (refer back to section 1.5.3) denoted by the symbol 'u', or equivalently, in energy units, commonly the electron volt or its multiples (such as the mega electron volt, MeV). Among these, the MeV is the more frequently encountered unit of mass and energy in nuclear literature. In atomic and molecular physics, on the other hand, the electron volt (eV) is the commonly used unit.

In this context, one has the conversion relations

$$1\text{u} = 1.66 \times 10^{-27} \text{ kg} = 931.5 \text{ MeV},$$

where the numbers quoted are approximate values.

The masses of the electron, the proton, and the neutron are as follows (once again the numbers represent approximate values):

$$\text{electron mass } m_e = 9.11 \times 10^{-31} \text{ kg} = 0.511 \text{ MeV},$$

$$\text{proton mass } m_p = 1.673 \times 10^{-27} \text{ kg} = 938.3 \text{ MeV},$$

$$\text{neutron mass } m_n = 1.675 \times 10^{-27} \text{ kg} = 939.6 \text{ MeV}.$$

Masses of other sub-atomic particles like the muon and the mesons (see sec. 18.8.9.1) are also commonly expressed in MeV.

18.8.4.3 Relating nuclear mass to binding energy

With this background, one recognizes the expression in (18-2a) as the energy of the separated configuration, where m_p and m_n are now to be interpreted as the *rest* masses of the proton and the neutron respectively. In sec. 18.2, this was taken as the mass of the nucleus (and then, even of the entire atom) only in an approximate sense since the mass of the nucleus, i.e., of the bound configuration of the nucleons, differs from this expression by the *binding energy*, to which we now turn.

Starting from the bound configuration of the nucleons, one can imagine a process where each of the nucleons is torn away from the nucleus and is made to move away to an infinitely large distance in such a manner that its kinetic energy is zero when at a large distance from the rest of the nucleons. This process of tearing the nucleons away from one another will require a certain energy for each of the nucleons since the binding forces between the nucleons will have to be overcome in the process. Let the total energy required in this process of arriving at the separated configuration of the nucleons be E_B , this being the minimum energy required to move the nucleons from the bound to the unbound configuration (a larger amount of energy would result in the nucleons having non-zero kinetic energy in the separated state).

If the rest mass of the nucleus under consideration, with Z number of protons and $A - Z$ number of neutrons in it be denoted by $M(A, Z)$, so that its rest energy is $M(A, Z)c^2$ then, in accordance with the principle of energy conservation, the energy of the nucleons in the separated configuration has to be $M(A, Z)c^2 + E_B$, and this must then be the same as the expression (18-2a) expressed in energy units, i.e.,

$$M(A, Z)c^2 + E_B = Zm_p c^2 + (A - Z)m_n c^2. \quad (18-14)$$

This relation gives a quantitative expression for the nuclear binding energy in terms of the nuclear mass $M(A, Z)$ and the total rest energy of the nucleons in the separated configuration. At the same time, it gives us a way to calculate the nuclear mass from the rest masses of the nucleons, once an independent estimate of the binding energy is arrived at, because the nuclear mass is less than the sum of the separated rest masses

precisely by the binding energy (expressed in mass units).

18.8.4.4 Binding energy and nuclear stability

The binding energy is, in a sense, a measure of the stability of the nucleus. The greater the binding energy, the more it takes to tear the nucleons away from one another. However, a more appropriate quantity in this context is the *specific binding energy* rather than the binding energy itself where the specific binding energy is defined as the binding energy *per nucleon*:

$$B_s = \frac{E_B}{A}. \quad (18-15)$$

The stability of the nucleus is, however, a concept that has several sides to it. Looking at the nucleus as a certain configuration (call it C) of the nucleons, with a certain mass-energy associated with it (equivalent to its rest mass), one may consider some *other* configuration (say, C') of the same nucleons, with some mass-energy associated with *this* configuration as well, and then imagine a *process* by which the nucleons change over from the first to the second configuration.

For instance, a nucleus $[A, Z]$ (a short-hand notation to denote the nucleus as a bound system made up of Z number of protons and $(A-Z)$ number of neutrons; this is the configuration C in the present instance) may emit an alpha particle $[4, 2]$ in this notation; see sec. 18.8.7.1) so as to get converted into the nucleus $[A - 4, Z - 2]$. The configuration C' is then the combination of $[A - 4, Z - 2]$ and $[4, 2]$. Some nuclei are found to spontaneously undergo this process of *alpha decay*, i.e., they are *unstable* against alpha decay. An experimentally measurable indicator of how stable or unstable the nucleus is against the process of alpha decay is the *half life* associated with the decay process: the larger the half life, the more stable the nucleus is against the decay.

The possible processes whereby a certain configuration of nucleons may get changed to another configuration may be quite numerous and diverse. Such a process may, for instance, involve, in addition to the nucleus $[A, Z]$, some other particle which may be another nucleus (say, $[a, z]$) or even a particle other than a nucleon (such as a photon).

It then makes sense to talk of the energy cost for the process whereby the initial configuration, made of $[A, Z]$ and the other particle, changes over to a final configuration, where the latter may once again involve one or more nuclei along with other particles like, say, a photon, or a beta particle (see sec. 18.8.7.2).

Every such process involving the nucleus $[A, Z]$ has an energy cost associated with it and this energy cost gives a measure of the stability of the nucleus against the process under consideration.

Such a measure of the stability involves, in general, the binding energy of the nucleus under consideration. However, apart from the binding energy of the nucleus $[A, Z]$, other parameters like the binding energies of the other particles involved in the process under consideration are also required so as to arrive at the energy cost of the process. Still, a comparatively larger value of the binding energy of $[A, Z]$ leads to a larger energy cost while conversely, a smaller value of the binding energy leads to a lower energy cost. In this sense, the binding energy E_B (or, more appropriately, the specific binding energy B_s) is a *general* measure of stability.

18.8.5 The binding energy curve

The general problem of estimating the nuclear binding energy is a complex one. An accurate theoretical estimation requires a detailed knowledge of the pairwise interaction between the nucleons, and even then, the problem remains intractable because the nucleus is, in general, not just an assembly of two nucleons but a *many-body system* involving several nucleons interacting together (the notable exception is the deuteron $([2, 1])$, the calculation of the binding energy of which is one of the cornerstones in nuclear theory).

18.8.5.1 The starting point

A clever way to approach this complex problem is to start from a simple *model* of the nucleus and then to bring in refinements and modifications, incorporating more and more detailed aspects of the many-body problem in steps, thereby making better and

better approximations to the binding energy E_B . The formula for the binding energy that one derives at any stage of the process may involve certain parameters taken in from experimental observations.

The *simplest* model to start with is the one where the nucleus is likened to a liquid drop, based on the saturation property of nuclear forces. Ignoring the nucleons near the surface of the nucleus, all the nucleons can be assumed to be *equally bound* in the nucleus. This means, in particular, that the energy required to tear away any one nucleon from the nucleus is assumed to be the same as the energy necessary to tear away any other nucleon instead. In other words, the specific binding energy in this simple model has to be the same for *all* nuclei. Though this is not borne out universally by experimental observations, appreciable deviations from this prediction are found mostly for relatively *light* nuclei and, to a lesser extent for the relatively *heavy* ones as well. In between, a large number of moderately heavy nuclei, neither too light nor too heavy, are found to conform to the prediction of a *constant specific binding energy* for all nuclei.

Denoting this constant value of the specific binding energy by b , one arrives at the following simple formula for the binding energy of a nucleus $[A, Z]$:

$$E_B = bA. \tag{18-16}$$

Here the parameter b is to be determined empirically, from observations involving the masses of nuclei of moderate size. However, as the formula is modified to relate more closely to actual values of nuclear binding energies and masses, a different value will have to be chosen for b such that the modified formula best fits the observed data for *all* nuclei at the same time.

Two of the most important modifications that need to be introduced in the above first-guess formula are the ones relating to the *finite size* of the nucleus and to the *nuclear charge*.

18.8.5.2 Finite size effect: the surface correction

As I mentioned in sec. 18.8.3.1, the nucleons near the surface of a nucleus are bound to a lesser extent compared to those in the interior since these surface nucleons do not have other nucleons interacting with it from all sides, i.e., have some of their bonds unsaturated (recall that the concept of bonds here is nothing more than a suggestive way of describing the interactions between nucleons, borrowed from the context of molecular bonds (see section 18.9.1)).

In other words, the expression (18-16), where all the nucleons have been assumed to be saturated for the sake of simplicity, is actually an overestimate for the binding energy, and a more accurate expression has to include a correction term to be subtracted from it. This correction term has to be proportional to the number of surface nucleons i.e., to the surface area of the nucleus. Using the expression (18-11) for the nuclear radius, this correction term assumes the form $sA^{\frac{2}{3}}$ which means that, with the surface correction included, the expression (18-16) for the binding energy of the nucleus $[A, Z]$ is to be modified to

$$E_B = bA - sA^{\frac{2}{3}}. \quad (18-17)$$

In this expression, s is a constant like b that is to be determined by making use of empirical data.

Incidentally, the analogy between a nucleus and a liquid drop becomes more transparent when one looks at the energy accounting in this manner, since a liquid drop, or any mass of liquid, is also characterized by a surface energy due to the fact that the surface molecules of the liquid have a different degree of binding as compared to the molecules located in the interior (recall, in this context, our discussions in section 7.6 relating to surface energy and surface tension).

18.8.5.3 The effect of the nuclear charge

The formula for the binding energy needs modifications on other counts as well, one of these being the effect of the nuclear *charge*. In addition to the strong interaction forces binding the nucleons together, the protons in the nucleus exert a repulsive Coulomb force on one another, and this requires a further reduction in the expression for the binding energy. In order to estimate the necessary correction on this count, the nucleus can be assumed to be a charged sphere with a charge Ze , where the radius of the sphere is given by the expression (18-11). The electrical energy of this charged sphere can be worked out from the principles of electrostatics, and resulting expression is found to be proportional to $\frac{(Ze)^2}{R}$ (see sec. 11.9.3), where the charge is assumed to be uniformly and continuously distributed through the volume of the nucleus. If one takes into account the discreteness of the protons carrying the nuclear charge, then the factor Z^2 in this expression gets modified to $Z(Z - 1)$, and one then arrives at the following modified formula for the binding energy:

$$E_B = bA - sA^{\frac{2}{3}} - c\frac{Z(Z - 1)}{A^{\frac{1}{3}}}, \quad (18-18)$$

where c is yet another constant to be determined from empirical data, by fitting the formula for the binding energy to experimental mass data for a number of nuclei.

18.8.5.4 Other corrections: the mass formula

The surface energy and the Coulomb repulsion energy are not the only sources of correction in the binding energy formula. There are other correction terms of importance in the binding energy formula whose theoretical explanation I will have to skip in this introductory exposition. One of these is the *asymmetry energy* term.

It might appear from the formula (18-18) that, the fewer the number of protons in a nucleus, the larger will the binding energy be, resulting in a more stable nucleus. However, this leads to a large *asymmetry* between the number of protons and that of

neutrons in the nucleus which, in turn, leads to a reduction in the binding energy, tending to make the nucleus more unstable. In other words, one has to introduce an asymmetry energy term in the expression for the binding energy, which follows from quantum principles applied to the assembly of nucleons in the nucleus.

Finally, one has to include a *pairing term* in the expression for the binding energy, which arises from the pairing interaction between nucleons mentioned in sec. 18.8.3.

With all these corrections included, one ends up with a formula for the binding energy, and hence for the nuclear mass (by eq. (18-14)), referred to as the *semi-empirical mass formula* where the term 'semi-empirical' means that the formula, though derived from theoretical considerations, needs as input the values of a number of constants (such as b , s , and c above) to be meaningful. This we write as

$$E_B = bA - sA^{\frac{2}{3}} - c\frac{Z(Z-1)}{A^{\frac{1}{3}}} - a\frac{(A-2Z)^2}{A} - \sigma_P p A^{-\frac{3}{4}}. \quad (18-19)$$

In this formula, the fourth term on the right hand side corresponds to the asymmetry energy and the fifth term arises due to the pairing interaction. These two terms include the constants a and p respectively like the constants b , s , and c introduced above, and are to be determined by fitting the formula to empirical mass data. Finally, σ_P is the *pairing number* that can assume the values -1 , 0 , or $+1$ because the pairing term in the binding energy depends on whether there is any *unpaired* neutron or proton in the nucleus. Its value is -1 for *even-even* nuclei, i.e., nuclei for which both $N(\equiv A - Z)$ and Z are even, 0 for *even-odd* (N even, Z odd) or *odd-even* (N odd, Z even), and $+1$ for *odd-odd* (N odd, Z odd) nuclei.

With all these corrections to the binding energy included, one finally arrives at the semi-empirical mass formula (refer to relation (18-14))

$$M(A, Z) = Zm_p + (A - Z)m_n - \frac{E_B}{c^2}. \quad (18-20)$$

Evidently, an increase in the binding energy goes to making the nucleus relatively more stable and conversely, a lower value of the binding energy corresponds to the nucleus

being relatively less stable.

18.8.5.5 The graph

Making use of formula (18-20) with appropriate values of the empirical constants put in, one can draw a graph with the specific binding energy $B_s = \frac{E_B}{A}$ plotted against A . According to the formula, E_B depends on *both* Z and A . It may so happen that, for a given A , there occur several nuclei with different values of Z . This is related to the fact that for any given Z (it is the atomic number Z that gives identity to an element), there may occur nuclei with various different neutron numbers (termed *isotopes*), i.e., different values of A .

Among the nuclei with a given A and various different values of Z , there usually occurs one particular value for which the nucleus is a stable one, the nuclei with the other values of Z being unstable against *beta decay* (see sec. 18.8.7.2) (there are only a few known instances where there occur two stable nuclei with a given A and with two different values of Z). One chooses the stable nuclei for the various possible values of A in plotting the graph.

In reality, one needs *three* different graphs here because of the three different values of the pairing number σ_P and even then, the experimentally determined masses do not all correspond to points lying on the graph. Fig. 18-8 shows schematically the general nature, or the *trend* of variation of B_s with A . One observes that there occurs an approximate *plateau* region in the graph for nuclei whose mass numbers are neither too small nor too large (roughly, for $20 < A < 150$), corresponding to an almost constant value of the specific binding energy, for which a formula of the form (18-16) adequately explains the variation of the binding energy with A . This constant value of the specific binding energy is found to be somewhere around 8 MeV per nucleon. Here I have made use of the unit MeV, which stands for a million electron volt of energy (recall that $1 \text{ eV} = 1.6 \times 10^{-19} \text{ J}$), the commonly used unit of energy in nuclear physics.

The sharply rising part of the graph in fig. 18-8 for the comparatively light nuclei corresponds to the effect of the surface term in the binding energy. For a nucleus belonging

to this group, there is a preponderance of surface nucleons, causing the binding energy to be significantly less than the value given by a formula of the form (18-16). Finally, the relatively slowly falling part for heavy nuclei (for $A \sim 60$ onward) corresponds, in the main, to the effect of the Coulomb term in the formula (18-19). For these heavy nuclei, the proton number Z has a large value and the mutual repulsion between the protons causes the binding energy to decrease for increasing values of A .

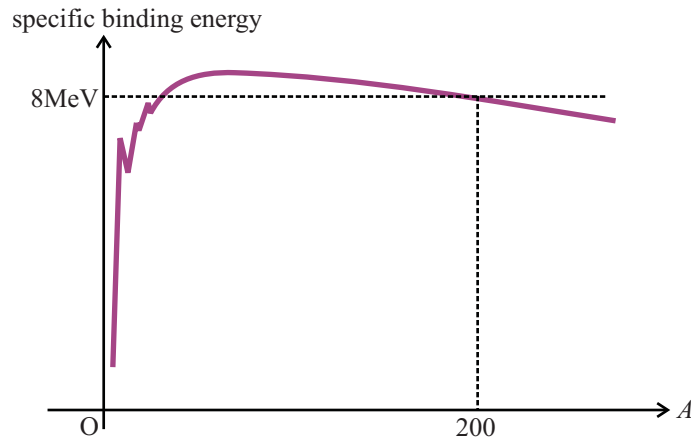


Figure 18-8: Binding energy graph; trend of variation of $\frac{E_B}{A}$ with A for stable nuclei.

Problem 18-5

Calculate the energy (in MeV) necessary to split an alpha particle into (a) two deuterons, and (b) two protons and two neutrons, given the following data: mass of a helium-4 atom, 4.00260 u; mass of a deuterium atom, 2.01410 u; mass of a hydrogen atom, 1.00783 u; mass of a neutron, 1.00867 u.

Answer to Problem 18-5

HINT: Suppose a parent nucleus A , with atomic number Z and nuclear mass m_A , is split into a number of daughter nuclei A_1, A_2, \dots , with atomic numbers Z_1, Z_2, \dots (hence $\sum_i Z_i = Z$) and nuclear masses m_{A_1}, m_{A_2}, \dots , where the daughter nuclei are all assumed to be at rest and separated from one another by a large distance. The energy required for this is given by $E = ((m_{A_1} + m_{A_2} + \dots) - m_A)c^2$. This can be expressed in the alternative form $E = ((M_{A_1} + M_{A_2} + \dots) - M_A)c^2$, where the upper case M 's denote *atomic* masses of the corresponding nuclear species.

This is so because the atomic mass (M_A) of A includes the mass of Z number of electrons while the total atomic mass ($M_{A_1} + M_{A_2} + \dots$) of A_1, A_2, \dots , also includes the mass of the same number of electrons. A small difference between the two expressions (one in terms of nuclear masses and the other in terms of the corresponding atomic masses) results from the binding energies of the electrons in respective atoms, but this can be ignored for most practical purposes.

(a) Thus, for an alpha particle to be split into two deuterons, one needs consider the energy equivalent of the mass difference between two deuterium atoms and a helium-4 atom, i.e., $E = 0.02560 \times 931.5 \text{ MeV}$, i.e., 23.85 MeV.

(b) In the case of an alpha particle being split into two protons and two neutrons, one has to consider the mass difference between a combination of two hydrogen atoms and two neutrons, and a helium-4 atom. Converting from atomic mass unit to MeV (multiplication by 931.5) one gets the required energy as 28.32 MeV. In other words, it takes less energy to split an alpha particle into two deuterons than to split it into two protons and two neutrons (as expected!).

18.8.6 Single particle and collective nuclear excitations

If an amount of energy is supplied to a nucleus by means of collisions with other particles, the energy so supplied may cause one or more nucleons to be knocked out from the nucleus by means of a *nuclear reaction*. Alternatively, the nucleus may be *excited* to a state of higher energy while the numbers of neutrons and protons in it remain unchanged.

A nucleus is an assembly of neutrons and protons obeying the basic rules of quantum theory, and may be in any one of its possible *stationary states*, where each stationary state is characterized by some definite value of energy. The stationary state with the lowest energy is termed the *ground state* while those with higher energies are termed the *excited states*. Under ordinary circumstances, the nucleus tends to remain in the ground state. If, by some means it is raised to an excited state then a process of *de-excitation* sets in whereby the energy of excitation is given out, usually in the form of photons.

The existence of excited states of a nucleus and the tendency of the nucleus to give away the energy of excitation so as to return to the ground state are similar to what one finds in the case of an atom. Recall that an atom can also be either in its ground state or in one of several possible excited states, and that an excited atom tends to return to its ground state by emitting photons. The energy of excitation and de-excitation of an atom is typically of the order of several electron volt while, on the other hand, nuclear excitation energies are typically of the order of several MeV (million electron volt). This comparatively large energy associated with nuclear excitations is responsible for a greater variety of processes that may possibly occur during de-excitation, some of which we will encounter below.

In summary, a nucleus, as an assembly of protons and neutrons, is a complex dynamical system subject to the rules of quantum theory, and may be in any one of a large number of possible stationary states. Under ordinary circumstances, it tends to reside in the ground state, while a supply of an amount of energy causes it to make a transition to an excited state. Subsequently, a process of de-excitation ensues, and the possible de-excitation processes may be quite varied and numerous. Among these, the one involving the emission of photons causes the excited nucleus to return to its ground state without changing its identity, i.e., keeping the values of A and Z unchanged.

Since the energy of the assembly of neutrons and protons increases in nuclear excitation, there also occurs an increase in the nuclear mass as a consequence of the mass-energy equivalence principle. While talking of nuclear masses, as for instance, in the formula (18-20), one normally refers to the mass of the nucleus in its ground state. The mass of the excited nucleus is then larger than the ground state mass by the mass equivalent of the excitation energy. Equivalently, the rest energy of the excited state is larger than that of the ground state by the excitation energy.

The exact description of the excited states of a nucleus is not an easy job. However, one can broadly classify the excited states in a few groups. One of these consists of states that can be described as *single-particle* excitations, where the excitation energy goes to just one of the nucleons, leaving the rest undisturbed, raising that particular nucleon

to one of its higher single-particle states. This is similar to an electron in an atom being raised to one of the single-electron excited states while the other electrons remain in their respective ground states.

Considering any single nucleon in a nucleus, one can look at its motion in the field of force exerted by all the other nucleons taken together, arriving thereby at a description of the single-nucleon ground state and excited states. The ground state of the assembly of nucleons may then be described by following a procedure analogous to the *aufbau* principle for the electrons in an atom. Once the ground state of the nucleus is known, one can then consider states of the nucleus where any one of the nucleons is raised to a higher single-nucleon state. In addition, excited states where two nucleons are independently excited to higher single nucleon states may also be considered.

This approach of describing nuclear states in terms of excitations of individual nucleons is referred to as the *single-particle model of nuclear excitations*. It is found to lead to an acceptable description of the *low-lying* excited states of the nucleus, i.e., states with energies only slightly higher than the ground state energy.

By contrast, if the nucleus is excited by supplying a comparatively large amount of energy to it, then this energy is taken up by *all* the nucleons simultaneously, and the state of motion of the nucleus *as a whole* gets altered. For instance, the nucleus may start *rotating* like a rigid body about some axis, or there may occur *oscillations in the shape* of the nucleus similar to those of a liquid drop where the drop alternately assumes an elongated shape and then returns to a nearly spherical shape periodically. Such excitations of the nucleus are referred to as its *collective* modes.

In *summary* again, when a nucleus receives a supply of energy by means of, say, a collision with some other particle, it may get raised to an excited state where the excitation may be described, depending on circumstances, either as a single-particle or as a collective mode (modes of an intermediate description are also possible, and are more difficult to describe). The nucleus may subsequently return to its ground state by the emission of one or more photons. Alternatively, on interacting with the particle from

which it receives the supply of energy, the nucleus may give out one or more nucleons so as to change identity, thereby going over to a new configuration with altered neutron and proton numbers - a process commonly described as a nuclear reaction.

The low-lying excited states of a nucleus whose energies are close to the ground state energy can be represented in an *energy level diagram* in a manner analogous to the one depicting the energy levels of an atom. A number of nuclear processes involving the nucleus under consideration can then be conveniently described by referring to such an energy level diagram. Nuclear scientists have, over decades, amassed a stupendous amount of experimental and theoretical data giving the energy level diagrams of known nuclei.

18.8.7 Radioactive decay

A considerable number of nuclei occurring in substances found in nature and of those produced artificially by means of nuclear processes of various types, are found to undergo the process of *radioactive decay* whereby they spontaneously give out any one or more of three specific types of particles, these being respectively the *alpha particle* (see sec. 18.8.7.1), the *beta particle* (see sec. 18.8.7.2), and the *photon* (also referred to as the *gamma particle*). Among these, the emission of photons (gamma emission, or gamma decay) is simply the process where an excited nucleus returns to its ground state, while the other two involve a change in the identity of the nucleus, thereby being special instance of a broad class of processes referred to as *nuclear transmutations*.

Radioactivity is the process whereby a given assembly of neutrons and protons makes a transition from an unstable configuration to a more stable one. While a large number of naturally occurring and artificially produced nuclei are known, only a relatively small number (about 10 per cent of all) are found to be stable.

The relatively heavier of these nuclei are found to undergo alpha decay while radioactivity by beta decay occurs for all mass numbers, in each case tending to select out the most stable nucleus among a set of isobars (nuclei with a fixed A , but with differing values of Z). As mentioned above, gamma emission is a process of a different kind

where a nucleus tends to move over from an excited state to its ground state without having its identity changed. As we will see below, gamma decay is a process that often accompanies an alpha decay or a beta decay.

A radioactive decay process can be expressed symbolically in the form

$$A \rightarrow B + b, \quad [1]$$

where A stands for the ‘parent’ nucleus undergoing the decay, B for the ‘daughter’ nucleus resulting from the decay, and b stands for an alpha particle, a beta particle, or a photon, as the case may be. In the last case, A and B correspond to the same nucleus, with A being in an excited state at a higher energy compared to B.

This process represents a change in the configuration of an assembly of neutrons and protons from a less stable to a relatively more stable configuration, if the rest mass of A is larger than the rest masses of B and b taken together where, in the case of gamma decay, the rest mass of the photon is to be taken as zero. This gives the following condition for the process [1] to occur spontaneously:

$$m_A > m_B + m_b. \quad (18-21)$$

The difference of the rest mass energies of the configurations on the left and right hand sides of [1] appears as kinetic energies of B and b, where the requirement of the conservation of momentum implies that the kinetic energy of b is usually much larger than that of B because of the relatively large mass of the latter.

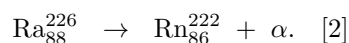
18.8.7.1 Alpha decay

The alpha particle is an assembly of two protons and two neutrons, being identical to a helium nucleus and is represented by the symbol α or He_2^4 where the numerals 4 and 2 are indicative of the mass number and the atomic number respectively (thus, the symbol X_Z^A stands for the nucleus of an element with chemical symbol X, and with mass number and atomic number A and Z respectively; an alternative symbolic representation of the

same nucleus would be $[A, Z]$).

The reason why a large number of nuclei ‘prefer’ to emit an alpha particle so as to go over to a relatively more stable configuration is that *the alpha particle is an exceptionally stable bound configuration of nucleons*. Since it contains 2 neutrons and 2 protons, the asymmetry term in the expression (18-19) for the binding energy is zero (recall that an increase in the value of the binding energy implies a reduction in the nuclear mass and a corresponding enhancement in the stability of the nucleus, which means that a smaller value of the asymmetry term corresponds to a more stable nucleus). In addition, the alpha particle is an even-even nucleus, for which the pairing term gives a positive contribution to the binding energy, i.e., corresponds to a higher degree of nuclear stability. This makes for a considerably high likelihood for the condition (18-21) to be satisfied when b is an alpha particle. Indeed, the condition (18-21) can be interpreted as corresponding to the requirement that the average binding energy per nucleon in the initial configuration has to be less than that in the final configuration. This is precisely what is likely to happen if an alpha particle is there in the final configuration.

Here is an example of alpha decay:



In this decay process a radium nucleus (‘parent’) emits an alpha particle to get converted into a radon nucleus (‘daughter’). A sample of radium, when kept in a closed container, gets gradually converted into radon and helium gases which collect in the container, while the amount of unconverted radium goes on diminishing. Recall that the alpha particle is identical with a helium nucleus, and each emitted alpha particle collects two electrons from its surroundings to give a helium atom. The total number of electrons in a radium atom being eighty eight, the daughter nucleus (radon) retains eighty six of these to form a radon atom while the remaining two are collected by the alpha particle.

Here is a simplified but useful picture describing the mechanism underlying alpha decay. The neutrons and the protons in the parent nucleus are in a constant state of

motion, during which clusters made up of two protons and two neutrons get formed in occurrences of a statistical nature. Once formed, such a cluster retains its identity for some time since the combination of two protons and two neutrons is an exceptionally stable one. Assuming that the condition (18-21) is satisfied, a cluster of this type is likely to come out of the nucleus as it moves about within the latter and hits its boundary. What prevents the cluster to actually come out is the strong attraction it feels due to the remaining nucleons in the nucleus. The situation is a bit tricky here since the breaking loose of the cluster from the rest of the nucleus is favoured in terms of energy balance, but the cluster is still to negotiate a barrier that it faces due to the strong attraction by the nucleons of the daughter nucleus.

The rules of quantum theory predict that, in such a situation, there is a finite probability that the cluster, on hitting the boundary of the nucleus, does succeed in overcoming the barrier, even though such a process is forbidden classically. This quantum mechanical process is termed *tunnelling*, or barrier penetration, and is once again statistical in nature. This means that every once in a while an alpha particle is given out by the nucleus under consideration with a certain probability, provided the energy condition (18-21) is satisfied. It is this probability that determines the *rate* of radioactive decay of a sample made of a large number of nuclei of the parent nucleus, and the *half life* of the latter, where the term 'half life' will be explained in sec. 18.8.7.4.

18.8.7.2 Beta decay

The beta particle comes in two varieties, the negatively charged beta (β^-) or the *electron* (commonly denoted as e^-), and the positively charge beta (β^+) or the *positron* (e^+). One difference, though, between the negatively charged beta particles emitted in radioactive disintegration and the extranuclear electrons in the atoms is that the former are highly energetic compared to the latter and are generated from within the nuclei undergoing the process of beta decay.

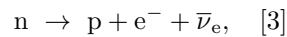
Corresponding to the two types of beta particles, there are possible two kinds of beta decay, namely the β^- -decay, and the β^+ -decay (also referred to as the positron decay).

In addition, there is possible a *third* kind of beta decay, namely the *electron capture* where an electron from an atomic orbital is absorbed into the nucleus.

The mechanism underlying the process of beta decay is more subtle than that responsible for alpha decay. Recall that neither the electron nor the positron is a constituent of the atomic nucleus, which is made of neutrons and protons. A clue to this mechanism is obtained from the fact that a *free* neutron, i.e., one not bound to other nucleons, is found to undergo β^- -decay with a certain probability (and correspondingly, a certain *half life*, see sec. 18.8.7.4).

The forces responsible for the splitting off of a β^- particle from a neutron are of a distinctly different nature compared to the strong interaction forces binding the nucleons in a nucleus, and are termed *weak interaction* forces. The weak interaction is one of the fundamental interactions in nature (the other fundamental interactions are the strong interaction, the electromagnetic interaction, and the gravitational interaction; see sec. 18.8.9.5).

The neutron beta decay is represented symbolically in the form

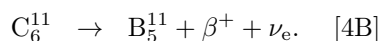
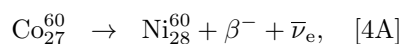


which shows that the neutron, on giving off an electron, gets converted into a proton and, in addition, *another* particle (denoted by $\bar{\nu}_e$) is given off in the process. A number of conservation principles, including the principles of conservation of energy and momentum, require the emission of this particle in the β^- -decay process. Its existence and involvement in the beta decay processes was predicted by Pauli on the basis of the principles of energy and momentum conservation, even before it was observed experimentally. This particle is called the *neutrino* (more precisely, the electron-type *anti*-neutrino). It is emitted in all nuclear β^- -decay processes. In a β^+ -decay process, on the other hand, the corresponding particle emitted is an electron-type neutrino, ν_e (the electron-type neutrino differs from two other varieties of neutrino; corresponding to all these three varieties, there exist their anti-particles).

The beta decay of the neutron is explained by making use of a number of basic features of the weak interaction forces, first suggested by Fermi in the context of beta decay of nuclei. The latter differs from the beta decay of the neutron in that it involves the conversion of a neutron into a proton in accordance with the scheme of the form [3], but the neutron, instead of being a free one, is *bound* to the other nucleons in the nucleus by strong interaction forces. Fermi's theory explains in quantitative terms a number of observed features of the process of beta decay, including ones relating to the *energy spectrum* of the beta particles. Since, in a beta decay process, the final configuration involves three particles - the daughter nucleus, the beta particle, and the neutrino (or the antineutrino, as the case may be), the energy released in the process can be shared by these particles in many different ways while at the same time conforming to the conservation of momentum, and hence the energies of the emitted beta particles are found to vary continuously over an interval ranging from zero up to a certain maximum energy. By contrast, in alpha emission, the final configuration involves just two particles, each released with a fixed kinetic energy in the process.

Fig. 18-9 depicts schematically a typical beta decay energy distribution graph, where the number of beta particles ($N(E)$) emitted per unit energy interval around each value of energy (E) of the beta particle is plotted against the energy.

Here are an example each of a β^- - and a β^+ -decay process.



The process of electron capture is a variant of the β^+ -decay where, instead of a positron being *emitted*, an electron from an extra-nuclear orbital is *absorbed* by the nucleus, along with the emission of a neutrino. An example of the electron capture process is given below

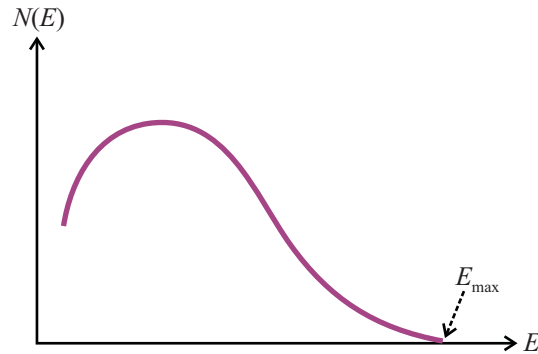
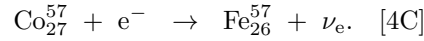


Figure 18-9: Beta decay spectrum, showing the variation of $N(E)$, the number of beta particles per unit energy interval around energy E , against the energy; the beta energy E is distributed continuously up to a certain maximum value E_{\max} where the latter is determined by the condition of the neutrino (or antineutrino, as the case may be) being emitted with zero energy.

Problem 18-6

Suppose a parent nucleus A_Z^A undergoes β^- decay, while at rest, into a daughter nucleus A'_{Z+1}^A along with an electron and an electron type antineutrino. If the atomic masses corresponding to the parent and daughter nuclei be respectively M_1 and M_2 in energy units, give an approximate expression for the maximum energy of the beta particles released in the decay of an ensemble of parent atoms.

Answer to Problem 18-6

Each parent nucleus breaks up into three particles - the daughter nucleus, the beta particle and the electron type antineutrino (which we refer to as the 'neutrino' for the sake of brevity) where one can assume the daughter nucleus to be much heavier than the beta particle, and the neutrino to be massless. The difference in mass-energy between the parent nucleus on the one hand and the daughter nucleus together with the beta particle on the other can be expressed as the mass difference (again, in energy units) between the parent and the daughter *atoms* (reason this out; the mass of a total of Z number of electrons gets canceled in calculating the mass difference), i.e., as $M_1 - M_2$.

This amount of energy appears as the kinetic energy of the three product particles. Of these, the kinetic energy of the daughter nucleus (commonly referred to as the recoil energy) can be ignored. This is because of the fact that the daughter nucleus carries a momentum of the same order in magnitude as the momenta of the beta particle and the neutrino (more precisely, the larger of the latter two) by virtue of the principle of conservation of momentum, but its energy is negligibly small because of its large mass (the energy E of a particle of mass M is related to the magnitude of its momentum (P) as $E = \frac{P^2}{2M}$; here we assume that the formulae of non-relativistic mechanics can be applied to the daughter nucleus). In other words, the energy $M_1 - M_2$ is shared between the beta particle and the neutrino.

The *maximum* beta energy (see fig. 18-9) E_{\max} corresponds to a particular instance of the decay process where the kinetic energy carried by the neutrino is zero. In other words, a reasonably good estimate for the maximum beta energy is $E_{\max} = (M_1 - M_2)$, where the right hand side is expressed in energy units.

18.8.7.3 Gamma decay

The process of gamma decay is represented by a reaction scheme of the form



where A^* stands for an excited state of the nucleus undergoing the decay, while A denotes a state of the same nucleus with a lower energy, commonly the *ground state* of the nucleus. Denoting the energy levels of the two states with the help of horizontal bars, with the energy increasing in the upward direction, a gamma decay process can be depicted diagrammatically as in fig. 18-10, where the wavy line running downward indicates the emission of a gamma photon. Since typical excitation energies of nuclei are of the order of an MeV, the emitted gamma photons are also of an energy of this order, which means that these are photons of frequencies far higher than those corresponding to visible light, and higher than even X-rays.

Gamma emission often accompanies alpha- or beta decay since in an alpha or a beta

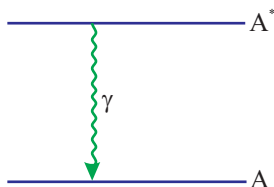


Figure 18-10: Excited state A^* and ground state A shown in an energy level diagram, in which a gamma emission is shown by a wavy line running downward.

decay, the daughter nucleus is commonly produced in an excited state. As the daughter nucleus returns to its ground state, it emits a gamma photon.

For instance, the energy level scheme for the process [2] above is depicted in fig. 18-11 where the two slanting lines indicate alpha emission from Ra_{88}^{226} , one to the ground state of Rn_{86}^{222} , and the other to an excited state of the same nucleus, the two energy levels of the daughter nucleus being drawn to the right of the energy level of the parent nucleus for the sake of clarity of presentation. For those decay events where the daughter nucleus is produced in the excited state, a gamma photon is emitted whereby the latter returns to the ground state.

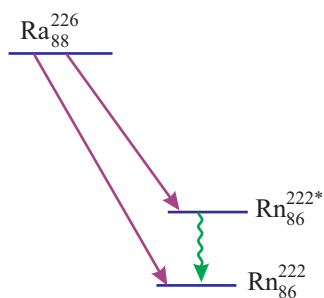


Figure 18-11: Alpha decay scheme of radium-226 showing gamma emission from excited state of the daughter nucleus; the daughter (radon-222) may be produced in either the ground state or an excited state (corresponding to distinct bunches of alpha particles produced in the decay); in the latter case, it gives off a gamma photon to descend to the ground state.

18.8.7.4 Radioactive decay law

Referring to a decay process of the form [1], suppose one starts with a sample made up of N_0 number of parent nuclei (A) at time $t = 0$. As the nuclei undergo radioactive

decay, the number of nuclei of the type A will decrease with time (with a corresponding increase in the number of daughter nuclei (B)). Let, after an interval of time t , the number of nuclei of type A surviving the process of disintegration be $N(t)$. In what relation does $N(t)$ stand to N_0 ?

An important feature of the radioactive decay process is that the disintegration of any one nucleus has *no correlation* with that of any *other* nucleus. Further, the disintegration of any one nucleus chosen at random, being a quantum mechanical process, is *statistical* in nature - one cannot make a prediction as to exactly when the nucleus will be disintegrating. Further, the fact that a nucleus has survived disintegration for a certain interval of time, has no bearing on how much more time it will take to disintegrate. The most specific statement that *can* be made relates to the *probability* of survival of any specified nucleus for a given interval t . making use of this probability, one can work out the *mean lifetime* of a nucleus in an assembly of parent nuclei. Every individual decay process, whether an alpha decay, a beta decay, or a gamma decay, is characterized by some specific value of the mean life, the inverse of which is commonly denoted by the symbol λ , where λ is referred to as the *disintegration constant*.

In terms of the disintegration constant λ , the relation between $N(t)$ and N_0 can be expressed as

$$N(t) = N_0 e^{-\lambda t}. \quad (18-22)$$

In other words, in a sample made of a large number of parent nuclei of a given type, the number of undisintegrated nuclei decreases exponentially with time. An alternative form of the relation (18-22) is the *differential equation*

$$\frac{dN}{dt} = -\lambda N, \quad (18-23)$$

i.e., the rate of disintegration at any given instant of time is proportional to the number of undisintegrated nuclei at that instant, the constant of proportionality being the inverse mean life λ . In the theory of nuclear disintegration, formula (18-23) is commonly

taken as the basic equation, starting from which one arrives at (18-22).

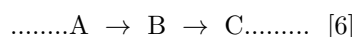
A constant related to λ is the *half life* of the substance (like, say, radium-226 in the case of the process [2] above) undergoing the decay. Starting with any given number (say, N) of the parent nucleus, the time interval required for this number to get reduced to half its value (i.e., to $\frac{N}{2}$) is defined as the half life. It is related to the disintegration constant as

$$\tau_{\text{half}} = \frac{\ln 2}{\ln \lambda}. \quad (18-24)$$

Equation (18-23), expressing the rate of disintegration for an assembly made of any given number of the parent nucleus with specified values of A and Z (a radioactive *species* as it is sometimes referred to), is known as the *law of radioactive disintegration*. However, a more fundamental view of the law relates to the probability of decay of a nucleus in any given time interval.

18.8.7.5 Successive radioactive disintegrations

Observations reveal the existence of *radioactive disintegration chains* where a chain is made up of a series of disintegrations as expressed by a decay scheme of the form



In this scheme, the parent species A is seen to decay into the daughter species B (by the emission of an alpha, a beta, or a gamma particle, which is not shown explicitly in the above representation) which, in turn, undergoes a disintegration into the species C , and the chain continues, finally ending up in a stable species that does not undergo further disintegration. The species A may itself be the product of some other radioactive species occurring earlier in the chain.

The rate of change of the number of nuclei of any given species in the chain can be worked out by invoking the law of radioactive disintegration on the assumption that the law operates independently for each disintegration. As a result, if N_A and N_B denote the

numbers of nuclei of species A and B respectively at any given instant of time, then one has

$$\frac{dN_B}{dt} = \lambda_A N_A - \lambda_B N_B. \quad (18-25)$$

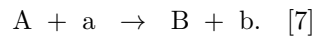
Here $\lambda_A N_A$ is the rate at which the number of B nuclei grows due to the disintegration of A (decay constant λ_A), while $\lambda_B N_B$ is the rate at which the number of B nuclei decays due to the disintegration of B (decay constant λ_B). The rates of change in the numbers of the other species of nuclei in the chain are also similarly obtained.

An example of a radioactive decay chain is the one where uranium-238 decays by alpha emission to thorium-234, followed by a further series of disintegrations, with the series terminating in lead-206.

18.8.8 Nuclear reactions

18.8.8.1 Introduction: examples of nuclear reactions

A radioactive disintegration by alpha, beta, or gamma emission is a spontaneous process in which the parent nucleus gets converted into the daughter nucleus all by itself, without any energy being given to it by means of a collision with any other particle. While belonging to the broad class of nuclear transmutations, a radioactive disintegration is distinguished from a nuclear *reaction* where the latter is typically a process of the following type

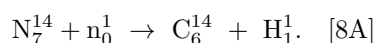


In this scheme, 'A' stands for a nucleus of a comparatively large value of the mass number and is termed the *target* nucleus, while 'a' is a comparatively light *projectile* nucleus. Experimentally, a fixed target made of species 'A' is hit by an energetic beam of species 'a'. There then takes place a nuclear transmutation giving rise to the species

'B' and 'b', where the mass number of 'B', the *residual* nucleus (also referred to as the daughter nucleus or the *recoil* nucleus), is typically large compared to that of 'b', and where the latter may, in special instances, even be a photon.

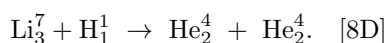
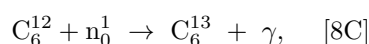
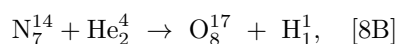
Nuclear reactions may be of quite numerous types, not all of which may conform to the above scheme. There may, for instance, be more than one light particles like 'b' emitted in the reaction. Or, there may be two daughter nuclei of roughly the same size produced, along with one or more light particles. Two important variants, namely the fission and fusion reactions, will be considered separately in sections 18.8.8.4 and 18.8.8.5.

As an example of a nuclear reaction, consider the process



In this reaction a neutron ($A = 1, Z = 0$) bombards a nitrogen-14 nucleus to produce a carbon-14 nucleus and a proton (the hydrogen nucleus). The reaction can be represented in the more compact form $\text{N}^{14}(\text{n}, \text{p})\text{C}^{14}$ (more generally, a reaction of the form [7] is represented as $\text{A}(\text{a}, \text{b})\text{B}$). This reaction is of considerable importance since it produces radioactive carbon by the bombardment of nitrogen nuclei in the atmosphere by atmospheric neutrons (these neutrons are produced by nuclear reactions occurring in the upper atmosphere). The radioactive carbon so produced helps in determining the age of fossils.

Here are a few other examples of nuclear reactions:



18.8.8.2 Conservation principles in nuclear reactions

Every nuclear reaction has to conform to a number of *conservation principles*. For instance, the total number of nucleons in the initial configuration (made up of the species 'A' and 'a' in the scheme [7]) has to be the same as that in the final configuration. Another conservation principle holds in respect of the amounts of charge in the initial and final configurations.

Other conservation principles of importance relate to the conservation of energy and momentum. Speaking of energy conservation in a nuclear reaction, it is useful to keep in mind that a nuclear reaction like, say, the one represented by the scheme [7] can, in general, be looked upon as an *inelastic collision* between the target nucleus 'A' and the projectile nucleus 'a', where there occurs a mutual conversion between the kinetic energies of the particles involved and their internal energies depending on the binding between the nucleons. As a consequence of this conversion, the rest masses of the particles taken together in the initial configuration may not be the same as that in the final configuration.

18.8.8.3 Energy balance in nuclear reactions

The difference between the total rest mass energy in the initial and final configurations has to be made up for by the difference between the kinetic energies in the two configuration. Thus, referring to a process of the type [7], the conservation of energy for the process implies the following relation

$$m_A c^2 + m_a c^2 + T_A + T_a = m_B c^2 + m_b c^2 + T_B + T_b, \quad (18-26)$$

where the T 's represent the kinetic energies of the respective particles, and the m 's their respective rest masses. In the case of a photon, the rest mass is to be taken as zero while T is to be replaced with the energy ($E = h\nu$) of the photon, which is related to the photon momentum (p) as $E = pc$. As regards the masses, one may use either the nuclear masses or, alternatively, the atomic masses of the respective species, where the relation involving the latter is obtained by adding the masses of the appropriate number

of electrons on both sides of the above equation (the case of positron emission, however, requires a separate consideration). This is possible since the binding energies associated with the electrons in the respective atoms are small compared to the energies relevant in the nuclear reactions.

The expression

$$Q = m_A c^2 + m_a c^2 - m_B c^2 - m_b c^2, \quad (18-27)$$

is referred to as the *Q-value* of the reaction. It represents the net energy released in the process, which appears as the excess of kinetic energy in the final configuration over that in the initial configuration. Reactions with $Q > 0$ (resp. $Q < 0$) are referred to as *exoergic* (resp. *endoergic*) ones.

In an endoergic reaction of the type [7], if the target nucleus 'A' be at rest in a given frame of reference (commonly the 'laboratory frame'), then a certain minimum energy T_0 of the projectile 'a' is needed to make the reaction occur. This is known as the *threshold energy* of the endoergic reaction under consideration.

Problem 18-7

A nucleus A_Z^A undergoes α -decay while at rest, according to the reaction scheme $A_Z^A \rightarrow B_{Z-2}^{A-4} + \alpha$. Find the kinetic energy of the α particle released in the decay in terms of the atomic masses of the three particles.

Answer to Problem 18-7

HINT: As explained above, the *Q*-value of the reaction is $Q = (m_A - m_B - m_\alpha)c^2$, where the notation is self-explanatory, and the m 's stand for the respective atomic masses. Since the nucleus 'A' decays from rest, this must be equal to the total kinetic energy of 'B' and the alpha particle: $Q = T_B + T_\alpha$. The two kinetic energies are related by the fact that the momentum of the daughter nucleus 'B' (commonly referred to as the 'recoil momentum') must be equal and opposite of the momentum of the alpha particle (in virtue of the principle of conservation of momentum), i.e., $\sqrt{2m_B T_B} = \sqrt{2m_\alpha T_\alpha}$. This gives $Q = T_\alpha + \frac{m_\alpha}{m_B} T_\alpha$, i.e., $T_\alpha = \frac{1}{1 + \frac{m_\alpha}{m_B}} (m_A - m_B - m_\alpha)c^2$.

Problem 18-8

If $A+a \rightarrow B+b$ represents an endoergic reaction, determine its threshold energy in terms of the atomic masses corresponding to the four nuclei.

Answer to Problem 18-8

HINT: The Q -value of the reaction, written in terms of the atomic masses, is given by $Q = m_A + m_a - m_B - m_b$, which is a negative quantity in virtue of the fact that the reaction is an endoergic one. On the face of it, it would seem as if the threshold energy should be just $-Q$, since the latter represents the energy deficit, in terms of the respective masses, of the initial particles with respect to the final products. However, this would imply that the final particles 'B' and 'b' would be produced at rest which, in turn, would be a violation of the principle of momentum conservation since the momentum of 'a' would then remain unbalanced.

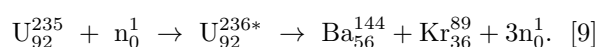
The minimum kinetic energy of 'a' necessary for the reaction is obtained by referring to the *center of mass frame* where, by definition, the momentum of the initial system of particles, *as also* of the final system, is zero. Hence, if we consider the case when the particles 'B' and 'b' are produced at rest in this frame, then that will correspond to the minimum kinetic energy necessary to make the reaction occur. Referring to this situation, then, let T , T' be the kinetic energies of 'a', 'A' in the center of mass frame, which have to satisfy $T + T' = -Q$ (ignoring the relativistic variation of the masses involved). But T and T' are to be related as $\sqrt{2m_a T} = \sqrt{2m_A T'}$ since the sum of momenta of 'a' and 'A' are to be zero (reason this out; the two particles must have equal and oppositely directed momenta), which implies $T' = \frac{m_a}{m_A} T$, i.e., $T = \frac{m_B + m_b - m_A - m_a}{1 + \frac{m_a}{m_A}}$.

Recalling that this is the required minimum energy of 'a' in the center of mass frame, the threshold energy (i.e., the minimum energy calculated in the laboratory frame, the latter being the frame in which the target particle 'A' is at rest) is obtained as $T_0 = (1 + \frac{m_a}{m_A})^2 T$ (check this out by making use of the definition of the center of mass frame), i.e., $T_0 = (1 + \frac{m_a}{m_A})(m_B + m_b - m_A - m_a)$.

18.8.8.4 Nuclear fission

A nuclear fission is a reaction where a heavy nucleus breaks up into two lighter nuclei of comparable size, with a few neutrons coming out in the process. A fission event is also commonly followed by a number of β^- emissions.

The fission process is usually initiated by a neutron being absorbed into heavy nucleus under consideration whereby an intermediate *compound nucleus* is formed in an unstable excited state. The instability gives rise to self-amplifying fluctuations in the nuclear configuration, finally culminating in the break-up of the nucleus. The following is an example of nuclear fission, where a uranium-235 nucleus absorbs a neutron, giving rise to an excited uranium-236 compound nucleus, subsequently breaking up into a barium-144 and a krypton-89 nucleus, giving off three neutrons in the process:



Here uranium-235 is the *fissile* nucleus while barium-144 and krypton-89 constitute the two *fission fragments*.

Depending on the fissile nucleus, the neutron triggering the fission process may be required to be an energetic one (a *fast* neutron) while, more importantly (and *unfortunately*), even a *slow* neutron may be capable of triggering the fission process, as in the above example of the fission of a uranium-235 nucleus.

What is unfortunate in the possibility of a fission being triggered by a slow neutron is that it makes a fission bomb work.

The possibility of a heavy nucleus breaking up into two fragments of comparable size rests on the shape of the binding energy curve of fig. 18-8. A heavy nucleus belonging to the slowly falling portion of the curve is characterized by a lower value of the binding energy per nucleon as compared to the nuclei of medium size in the middle of the curve where the binding energy per nucleon has a higher value. This is a consequence of the role of the Coulomb repulsion among the protons in a nucleus, which causes the binding energy to decrease with increasing Z (refer to the third term in eq. (18-19)). This makes the configuration of two fission fragments of comparable size a more stable one compared to the entire nuclear charge being concentrated in one single nucleus.

Even though the final configuration involving the two fission fragments may be charac-

terized by a lower energy as compared to the initial fissile nucleus, the latter may not be capable of undergoing the fission process spontaneously since the initial configuration may face a *barrier* which it has to negotiate before it can go over to the final configuration. The absorption of a neutron by the fissile nucleus gives it an extra energy, which helps the fissile nucleus negotiate the barrier and thus acts as a trigger for the fission to occur.

While the process of nuclear fission is triggered by the absorption of a neutron by a fissile nucleus, the fission process itself causes a number of neutrons to be released. This leads to the possibility of a *chain reaction* being set up in a material made up of fissile nuclei because the neutrons released in a fission event may be absorbed by other fissile nuclei in the material, thereby producing further fission events.

Chain reactions involving cascades of fission events are made use of in the generation of power in nuclear reactors (a process involving *radiation hazards*) as also, unfortunately, in the production of fission bombs.

Fission barrier.

In closing this section, I will include a few words on the *energy criterion* for nuclear fission.

Nuclear fission involves a competition between the *surface energy* and the *Coulomb energy* of a nucleus. Referring back to the binding energy curve introduced in sec. 18.8.5.5, one observes that the splitting of a parent nucleus into two daughter nuclei is accompanied by a decrease in the Coulomb energy of the constituent set of nucleons since the average separation of the protons gets larger in the process. On the other hand, looking at the parent and the daughters as liquid drops, the surface energy of the set of nucleons increases. For the nuclei in the slowly falling part of the binding energy curve, the process of splitting of the parent nucleus into a pair of daughter nuclei is energetically favorable in that the decrease in Coulomb energy then overrides the increase in surface energy. However, fission is a special instance of splitting where the two daughter nuclei

are of approximately same size (thus, the emission of an alpha particle from a heavy nucleus is cannot be termed a fission). For a fission process to be energetically possible, one has to move further along the binding energy curve toward heavier nuclei. It turns out that the fission of a heavy nucleus with atomic number Z and mass number A becomes energetically favorable if the criterion

$$\frac{Z^2}{A} > 18, \quad (18-28)$$

is satisfied, though it must be mentioned at the same time that this is not a hard and fast condition.

Recall that the Coulomb energy of a nucleus goes like $\frac{Z^2}{A^{\frac{1}{3}}}$ (refer to the third term on the right hand side of formula (18-18)), while the surface energy goes like $A^{\frac{2}{3}}$. The ratio $\frac{Z^2}{A}$ therefore expresses the relative importance of the Coulomb energy with respect to the surface energy.

Because of the mutual repulsion between the protons in a nucleus, the latter is not of a spherical shape even when the parameter $\frac{Z^2}{A}$ has a relatively low value, but is elongated to some extent, the degree of elongation increasing for nuclei with increasing values of $\frac{Z^2}{A}$. However, the elongated nucleus (call it N) can still be in an equilibrium configuration unless its mass-energy attains a value larger than combined mass-energy of daughter nuclei (call these A and B) in a separated configuration, assuming that N has gotten split into A and B by some means. For a nucleus satisfying the condition (18-28), this requirement is met with, which means that the separated A-B configuration (see fig. 18-12, where the nucleus is imagined to be like a deformable liquid drop) is energetically more favorable as compared to the bound configuration N.

However, in order that N may actually get split into the separated A-B configuration, it has to pass through intermediate stages where the elongation increases successively through configurations like N_1 , N_2 shown in the figure till a configuration N' is reached where there is a pronounced 'neck' or constriction which then gets severed so as to give

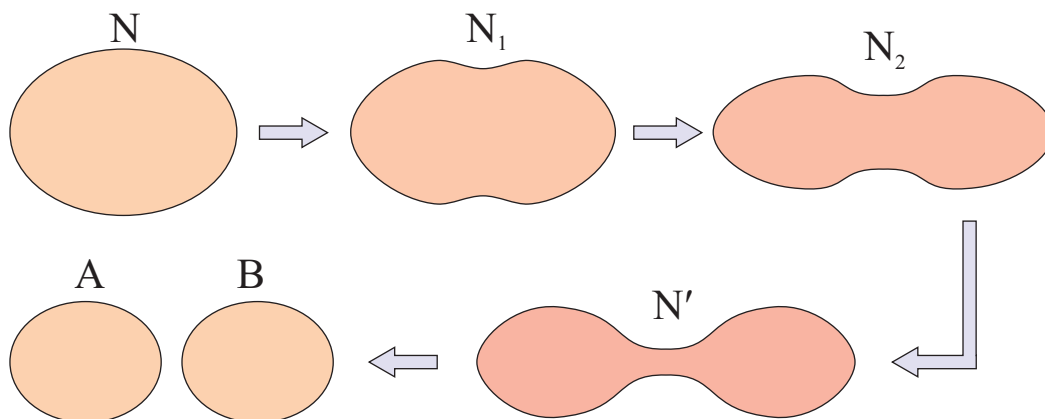


Figure 18-12: Successive configurations (schematic) through which a parent nucleus N has to pass so as to divide into daughter nuclei A and B separated from each other (configuration $A-B$ at the right); the nucleus gets more and more elongated (configurations N_1 , N_2) till a pronounced 'neck' is formed (configuration N') from which the separation takes place; the nucleus is imagined to behave like a deformable liquid drop and passes through the sequence of stages when excited with sufficient energy (see fig. 18-13(A)).

way to the configuration $A-B$. One can define a deformation parameter s that increases for the configurations shown in the figure from left to right, such that s has the value 0 for N , while the separated configuration $A-B$ (in which the daughters are at a large distance from each other) corresponds to a large value of s . One can then plot a graph with the energy as a function of the deformation s that looks like the one in fig. 18-13(A). One observes that the separated configuration ($s \rightarrow \infty$) has a lower energy as compared to the one with $s = 0$, as a result of which it is energetically favorable for N to get split into A and B , but there exists an *energy barrier* that must be crossed before the splitting can be actually possible. In other words, the nucleus N has to be excited by some means to the deformed configuration N' through which it has to pass for the fission process to be realized.

This excitation is commonly achieved by means of a slow neutron that is caused to be captured by N so as to get to the configuration N' ; in this case N' differs from N by an extra neutron count; however, the process of fission is usually accompanied with the shedding of a few neutrons from the nuclear mass, which makes the picture somewhat different from the simple representation in fig. 18-12 without, however, any major shift in the principle outlined.

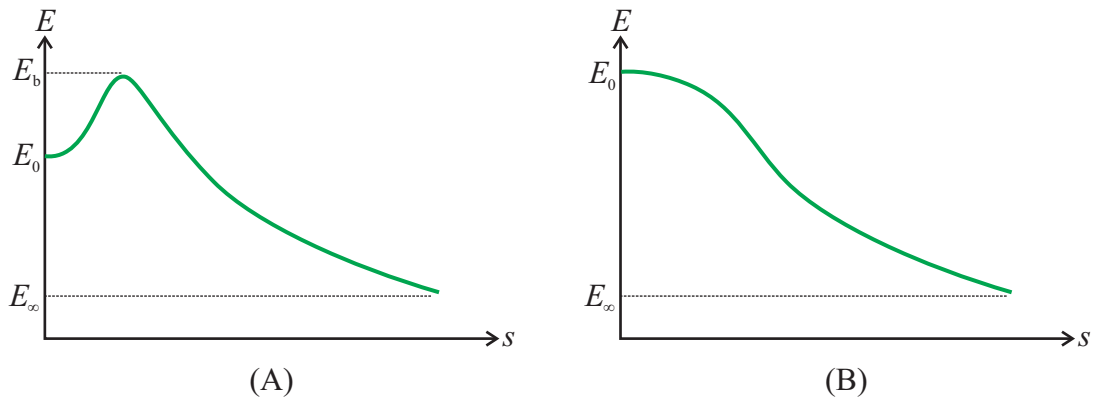


Figure 18-13: (A) Depicting schematically the energy barrier in fission for a nucleus satisfying the condition (18-28); referring to a simplified picture of the fission process depicted in fig. 18-12, the energy E of a nuclear configuration is plotted against the deformation parameter s ; the parent nucleus N ($s = 0$) corresponds to energy E_0 , which is higher than the energy (E_∞) of the separated configuration; there exists a fission barrier (energy $E_b > E_0$) that has to be crossed in order that the fission process can be realized; this corresponds to the configuration N' of fig. 18-12; once this configuration is crossed, the system moves downhill in the energy landscape; (B) for a nucleus satisfying (18-29), the barrier is non-existent and the nucleus is absolutely unstable against fission.

In other words, though the condition (18-28) tells us that the process of fission is energetically favored, still it does *not* imply *spontaneous* fission, since the nucleus still needs to be excited to make it cross the energy barrier. For nuclei with $\frac{Z^2}{A}$ values higher than the limit (recall that this limit is a rather loose one) indicated in the above condition, the barrier gets progressively reduced, till for nuclei with

$$\frac{Z^2}{A} \sim 47, \quad (18-29)$$

or higher it disappears altogether, as shown in fig. 18-13(B). Such nuclei are not found in nature since these are absolutely unstable against fission and get split spontaneously even if formed by some chance occurrence.

18.8.8.5 Nuclear fusion

Nuclear fusion is a process where two light nuclei are made to come together and get joined, or *fused*, into a single nucleus of larger size. Since the initial configuration is made up of two comparatively light nuclei, the average binding energy per nucleon in the

initial configuration corresponds to the rising portion of the graph shown schematically in fig. 18-8, with the specific binding energy plotted against the mass number. Since the final configuration corresponds to a nucleus with a larger value of A , the average binding energy per nucleon in the final configuration is higher, implying that the final configuration is a more stable one. In other words, the fusion process is an energetically favored one.

As explained in sec. 18.8.5.5, the initial rising portion of the specific binding energy curve is a consequence of the fact that the surface energy term in the formula (18-19) dominates for comparatively low values of A . Thus, fusion is the process by means of which the average surface energy per nucleon gets reduced.

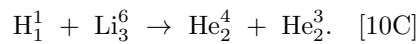
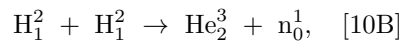
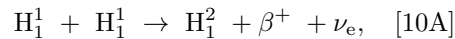
However, even though fusion of two comparatively light nuclei is an energetically favoured process, the two nuclei face a barrier in approaching each other sufficiently closely - a consequence of the Coulomb repulsion between the two nuclei, both of which are positively charged. In order that fusion may be possible, the nuclei are to be given a very high kinetic energy so that they may cross the barrier and get fused together. This, in turn, may be made possible by raising a gaseous sample containing the two nuclear species to a very high temperature. What is more, this may even set up a chain reaction of fusion events since the energy released in a fusion event may be utilized to maintain the gas at a high temperature, or even to raise its temperature to a still higher level, causing further fusion events to take place.

This approach of realizing a chain reaction of fusion events can lead to controlled power generation, but there are practical difficulties standing in the way. The heated gas is to be *confined* within a strictly limited volume so that it may not cause evaporation of the material of the containing vessel, and diffuse away. One way to effect this confinement is to use a magnetic field so that the Lorentz force on the charged nuclei (which no longer form neutral atoms by binding with electrons since all the electrons are knocked out of the atoms at such high temperatures, and the gas exists in a *plasma* state) make them move in coiled paths, thus confining the gas. However, such confinement is not of a *stable* nature - self-reinforcing oscillations of the gas in the plasma state make it

break away from the confined condition.

Here again, human ingenuity has not failed to follow a more diabolical course and has made possible the *uncontrolled* production of energy in the form of the fusion bomb.

Here are a few examples of fusion reactions:



Fusion reactions occurring in an environment of high temperature are sometimes referred to as *thermonuclear* reactions.

Thermonuclear reaction cycles are responsible for the energy generation in stars.

18.8.9 Introduction to elementary particles

According to the present state of knowledge, elementary particles and their interactions are described by the so-called *standard model*. There do remain a number of unresolved questions and the need for a broader theory, but the standard model is nevertheless a consistent theory capable of accounting for a wide range of experimental observations.

According to the standard model, elementary particles in nature belong to a number of families, and their interactions are explained in terms of a number of *exchange* particles. The interactions mediated by these exchange particles are also of a few fundamental types.

To begin with, the term elementary particles was used earlier to describe electrons, protons, and neutrons, the basic constituents of the atom and the atomic nucleus. How-

ever, experiments on cosmic rays in various strata of the atmosphere produced evidence of other particles that were as elementary as the protons, neutrons and the electrons. The development of the giant accelerators capable of producing particle beams of very high energy, and the resulting experiments using these high energy particles as projectiles further lengthened the list of elementary particles and, at the same time, indicated that most of these particles are endowed with some kind of internal structure.

On the theoretical side, the elementary particles and their interactions were sought to be explained in terms of *fields* that were thought to be the basic dynamical entities such that the particles in various states of motion could be interpreted as so many *quantum states* of the fields (see section 16.15 for a brief introduction). The quantum theory of the electromagnetic field was developed along these lines, registering impressive successes in resolving a large number of theoretical and practical questions. But a number of unresolved questions still remained and, moreover, the quantum theory of fields failed to accommodate satisfactorily the growing list of elementary particles.

This started the search for a more fundamental description of the particles and their interactions that resulted finally in the standard model. In the following, particles like the proton and the neutron will be included in the list of elementary particles, but these are not considered *fundamental* in the theory since these can be looked upon as composites of *quarks*, the latter being the fundamental entities in the standard model.

18.8.9.1 The classification of elementary particles

To start with, elementary particles are grouped into two broad classes, namely, the *leptons* and the *hadrons*, where the latter are further divided into two families - the *mesons* and the *baryons*.

The lepton family includes the electron (e^-), the *muon* (μ^-), and the *tau* particle (τ^-), along with *neutrinos* of three types, the electron neutrino (ν_e), the muon neutrino (ν_μ), and the tau neutrino (ν_τ). Along with these six particles, all their *anti-particles* are also included in the lepton family, where the concept of anti-particles will be introduced in sec. 18.8.9.3.

While the electron was known as an elementary constituent of matter from early days of atomic physics, the muon and the tau particle were observed in cosmic ray experiments. The existence of the electron neutrino was postulated by Pauli in explaining certain aspects of beta decay, while the other neutrinos are involved in the decay processes involving the muon and the tau particle.

The families of mesons and baryons include particles that are, in general, heavier than the leptons, but these differ from the leptons from a more fundamental point of view.

Each of the two families of mesons and baryons is further subdivided into a number of *multiplets*. Of these we will refer to only the *lowest* (i.e., the lightest) multiplet in each family. Each of the two lowest multiplets is made up of eight particles and is referred to as an *octet*. The particles making up the other, *higher* multiplets in each family are heavier than those in the lowest one.

The proton and the neutron belong to the lowest baryon octet while the *pi mesons* (π^\pm, π^0) observed in cosmic ray experiments belong to the lowest meson octet. The pi mesons were initially thought to be the *exchange* particles responsible for the binding of protons and neutrons in the nucleus. This idea of the mesons mediating the strong interactions of nucleons has later been supplemented by the concept of the nucleons themselves being bound systems of *quarks*, where the interaction among quarks, responsible for their binding is effected by means of a set of postulated particles called *gluons*.

The lowest baryon and meson octets include other particles as well, known as *strange* particles since they are characterized by a certain tell-tale feature termed *strangeness* (see sec. 18.8.9.2 below). These strange particles were also first discovered in cosmic ray experiments. Incidentally, the anti-particles of the baryons in the lowest octet form a separate octet among themselves, while the meson octet contains the mesons as well as their anti-particles (thus, for instance, π^- is the anti-particle of π^+).

This entire classification of elementary particles into leptons, mesons, and baryons is by no means an arbitrary one, since it reflects a number of fundamental principles

governing the world of elementary particles.

Finally, there are the *exchange* particles that act basically as mediators of interactions between the other elementary particles. The most commonly known exchange particle is the photon, the mediator of electromagnetic interactions, while there are other exchange particles as well, corresponding to the other fundamental interactions in nature.

18.8.9.2 Elementary particles and quantum numbers

Each elementary particle is characterized by a set of quantities specific to it, where the set of all these quantities taken together may be looked upon as the *definition* of the identity of the particle.

To begin with, each of the elementary particles can be either a *fermion* or a *boson*, where the characteristic of being a fermion or a boson is related to the *spin* of the particle.

The spin was introduced in section 18.3.3 while describing the various possible states of electrons in an atom, where it was identified as an *intrinsic* angular momentum. The spin of an elementary particle can be specified in terms of a certain *spin* quantum number, where the spin quantum number can be either an *integer* or a *half-integer*. For instance, the spin quantum number (or, simply, the *spin*) of the electron is $\frac{1}{2}$.

All particles with half-integer spins belong to the class of fermions, while those with integral spin are bosons. Apart from the spin, the distinction between fermions and bosons shows up in the description of states of *groups* of particles of any given kind. One manifestation of this distinction is the *Pauli exclusion principle* obeyed by fermions, but *not* by bosons (see sec. 18.4.1, 18.5.2.1, where the Pauli exclusion principle was introduced in the context of quantum states of electrons in an atom).

Apart from the spin and the fermionic or bosonic nature of a particle, there are other features defining its identity, of which the rest mass of the particle is one (at times, one quotes the rest mass *energy* instead of the rest mass). The rest mass differs from the spin in that the latter is specified in terms of a *quantum number* that can have any

one of a *discrete* set of values, while the rest mass can have any value belonging to a *continuous* range from zero to infinity.

Another defining property of a particle is its *charge*. While the charge, like the rest mass, does not correspond to a quantum number, it is, nevertheless, *quantized* in that the charge of any particle or a group of particles is an integral multiple of the basic quantum of charge, namely, the charge of an electron.

The charge of a quark (see sec. 18.8.9.4) can be a fraction of the electronic charge.

However, a quark is not observed independently as a single particle, but always occurs in combination with other quarks, the total charge of the combination being an integral multiple of the electronic charge.

A quantum number of fundamental relevance is the *intrinsic parity* of a particle, which relates to the way the description of the internal quantum state of the particle changes under the transformation from a right-handed to a left-handed co-ordinate system. The parity is a *multiplicative* quantum number that can have the value $+1$ or -1 for a particle or for a system made of several particles. Thus, if the intrinsic parity of a particle be P , then that of a system of N particles will be P^N . The proton, neutron, and the electron have intrinsic parity $+1$ each, assigned by convention.

The leptons are distinguished from other particles by means of a quantum number termed the *lepton number* (L), the lepton number of all these other particles being zero while that any of the six leptons is $+1$. Correspondingly, the lepton number of any of the anti-particles of these six leptons is -1 . However, while the lepton number distinguishes the leptons from other particles, there exists a finer distinction *within* the lepton family. Thus, for instance, the electron (e^-) and the electron neutrino (ν_e) have the *electron number* (L_e) $+1$ and their antiparticles (e^+ and $\bar{\nu}_e$) have $L_e = -1$, while all the other leptons have $L_e = 0$. The *muon number* and the *tau number* are similarly defined.

Analogously, the quantum number characterizing the baryon family is the *baryon number* (B), which is $+1$ for the baryons and -1 for the antibaryons.

Another quantum number characterizing the baryons and the mesons is *strangeness* (S). The proton and the neutron, the two lightest baryons have $S = 0$, and similarly the strangeness of the three lightest mesons (π^\pm , π^0) is zero, which is why all these particles are referred to as *non-strange* ones. A number of baryons and mesons of higher mass have $S \neq 0$ and are termed *strange* particles.

In this brief introduction, I do not enter into a description of the other quantum numbers relevant in the description of elementary particles and their interactions.

18.8.9.3 Anti-particles

To every elementary particle there corresponds an *anti-particle*, with the same mass and spin and with other characteristics *opposite* to those of the particle. Since the spin of a particle is the same as that of its anti-particle, either both are fermions or both are bosons. With reference to the other quantum numbers, a particle and its anti-particle are, in a sense, mirror images of each other. In some instances, a particle and its anti-particle are *identical* in all respects, i.e., the particle itself is its anti-particle, examples being the photon and the neutral pion (π^0).

All stable matter in our universe is composed of protons, neutrons, and electrons, while the production of any anti-particle results in a process of *annihilation* where the anti-particle and the corresponding particle are destroyed simultaneously, mostly giving rise to photons. Thus, anti-particles in our universe are short-lived objects, and *anti-matter*, made of anti-particles, does not occur in the form of stable materials. The *asymmetry* between matter and anti-matter is an interesting open question in particle physics as also in cosmology.

Examples of anti-particles are the anti-proton (\bar{p}) and the anti-neutron (\bar{n}). Anti-particles are often denoted by putting a short horizontal line on top of the symbol for the corresponding particle, while a separate symbol for the anti-particle is also used in some instances (e.g., e^+ as the anti-particle of e^-).

18.8.9.4 The quark structure of elementary particles

As I have mentioned above, particles like the proton, the neutron, and the mesons seem to have an internal structure of their own, implying that these are not truly of an elementary nature. On the other hand, such internal structures have not been discerned in the case of leptons. According to the current theory of elementary particles, the standard model, the mesons and baryons are made up of a set of more fundamental entities called *quarks*. A consistent theory explaining the structures of all the mesons and baryons observed so far requires the existence of six types of quarks, and these different types are referred to as quarks of different *flavors*. The six quarks are denoted by symbols u (*up* quark), d (*down* quark), s (*strange* quark), c (*charmed* quark), b (*beauty* quark), and t (*truth* quark). The quarks are all spin- $\frac{1}{2}$ fermions, and have *fractional* charges (respectively, $\frac{2}{3}, -\frac{1}{3}, -\frac{1}{3}, \frac{2}{3}, -\frac{1}{3},$ and $\frac{2}{3}$ for the above six quarks). Moreover, each of these quarks have a fractional baryon number ($\frac{1}{3}$) as well. Associated with each quark (q), there is an *anti-quark* with opposite quantum numbers. The six quarks and their anti-quarks form a family analogous to the six leptons and their anti-leptons.

The quarks are always found to occur in bound configurations, with a quark and an anti-quark forming a meson and three quarks forming a baryon, these combinations having integral values of charge and baryon number. For instance, the proton results from the combination (uud) and the neutron from (udd), while the π^\pm meson correspond to the combinations $u\bar{d}$ and $\bar{u}d$ respectively.

Some of these combinations require two quarks of the same flavor to be in the *same* quantum state, which goes against the Pauli exclusion principle. In order to overcome this problem, the standard model has postulated the existence of three varieties of quarks of each flavor, these three varieties being described as quarks of different *colors* - red, blue, and green. A baryon is a *colorless* combination of three quarks, there being one quark each of the three colors in the combination. Similarly, a meson is also colorless in that it contains a quark-anti-quark combination of the same flavor and color. I will, however, not enter into a more detailed consideration of the color characteristic of the quarks.

18.8.9.5 The basic interactions

The classification of elementary particles and the assignment of quantum numbers to them, or the assumption of the underlying quark structure, all are based on observations of various processes, where these processes are caused by *interactions* between the particles or, at a more fundamental level, between the building blocks they are made of.

All such interactions observed in nature have been found to be fundamentally of *four* types, each type having its own characteristic features and underlying mechanisms distinguishing it from the others. These are the *strong*, *electromagnetic*, *weak*, and *gravitational* interactions.

The gravitational interaction is the one responsible for the motions of celestial bodies under their mutual attraction, and of projectiles under the force of gravity. The basic law underlying these motions is Newton's law of gravitation introduced in chapter 5. However, as current observations go, this interaction has negligible observable effect at the microscopic level in elementary particle interactions since it is the weakest of all the above four types of interactions.

Gravitation, however, assumes fundamental relevance over extremely small length- and time scales where one feels the need to describe all the other fundamental interactions in a new setting. This, however, constitutes one of the basic open problems in physics.

In an ordering according to the strength of the interactions, the weak interaction comes next. It is the weak interaction that is responsible for the beta decay of nuclei and a number of other phenomena in the world of elementary particles. Like the strong interaction, it has no direct observational effect in the interaction of macroscopic bodies since these are of an extremely short range (of the order of a fraction of a fermi). The electromagnetic interaction has an intermediate strength between the weak and the strong interactions but, similar to the gravitational interaction, has a long range (recall

that electrical and gravitational forces both obey an inverse square law) and explains a wide range of observed phenomena at a microscopic *and* a macroscopic level. Finally, the strong interaction features at the top of the list based on the strength of various interactions, but is of a short range (of the order of a fermi). As we have seen, it is responsible for the binding and stability of the atomic nucleus.

On a comparative scale of strength the strong, electromagnetic, weak, and gravitational forces occur in a ratio of the order of $1 : 10^{-2} : 10^{-9} : 10^{-38}$.

Interestingly, the distinction between leptons on the one hand, and hadrons (i.e., mesons and baryons) on the other, relates to the types of interaction they can participate in. Thus, *leptons do not participate in strong interactions*, while hadrons participate in both strong and weak interactions (incidentally, phrases like strong interactions or weak interactions (plural) are commonly used to refer to *processes* governed by the strong or the weak interaction (singular)).

18.8.9.6 The conservation principles

Apart from the range and strength, the various interactions differ in the *conservation laws* obeyed in the elementary particle processes they lead to. Conservation of energy, momentum, angular momentum, and charge, are *absolute* in that these are found to hold good in *all* the interactions. Among the quantities described by discrete quantum numbers, the baryon number and the three lepton numbers are also conserved in all the interactions.

Numerous other quantum numbers, on the other hand, are *differentially* conserved in that their conservation holds in some of the interactions but not in others. For instance, *isospin* (refer to section 18.8.2 where the concept of isospin was introduced in the context of the distinction between the proton and the neutron) is conserved in strong interactions (resulting in the charge independence of the strong interaction) but not in electromagnetic and weak interactions. Similarly, parity is conserved in strong and electromagnetic interactions but not in weak interactions (processes governed by the weak interaction are not always mirror-symmetric).

Strangeness is a quantum number that is conserved in strong interactions but not in weak interactions. Strange mesons and baryons are produced in pairs in the upper atmosphere from non-strange hadrons by means of the strong interaction while the decay of a strange particle cannot occur through the strong interaction since such a process would violate the conservation of strangeness (this statement applies to a comparatively light strange particle where there does not exist a lighter strange particle it can decay into). Thus, a strange hadron decays to a lighter, non-strange hadron, by means of the weak interaction. An example of such a decay is the process

$$\Lambda^0 \rightarrow \pi^- + p \quad [11],$$

where Λ^0 is a strange baryon ($S = -1$) belonging to the lowest baryon octet.

18.8.9.7 The mediating particles

According to the classical view of nature, the interaction between two charged particles is effected by means of electromagnetic waves carrying energy, momentum, and angular momentum from one particle to the other. In the quantum view, various quantum states of the electromagnetic field are described in terms of *photons*, the quanta of the field, and electromagnetic interactions can then be thought of as being mediated by means of photons. In a similar manner, the strong and the weak interactions are also mediated by particles analogous to photons. All these mediators are *bosons* having integral spins, there being a characteristic set of mediating particles for each of these interactions.

For instance, the weak interactions are mediated by a group of three particles, the W^\pm and the Z^0 bosons. These are quite heavy particles and can be produced in experiments at very high energies.

There exists a neat theoretical description of the weak interactions mediated by the W^\pm and Z^0 bosons in terms of *diagrams*, where a diagram involves a set of *vertices*. Each vertex, in turn, is made up of a pair of lines corresponding to two leptons or to two quarks, and a third line corresponding to a mediator boson. Drawing out a diagram for a given process, one can work out the *probability* of occurrence of the process by

invoking a set of specific rules of calculation. At a deeper level, this approach of working out probabilities from the diagrams rests on the quantum theory of *fields*, where the elementary particles appear as *quantum states* of the underlying fields.

Electromagnetic interactions are similarly accounted for in terms of diagrams representing the possible processes involving the fields corresponding to the various charged particles and the electromagnetic field. A vertex in any such diagram involves a photon line instead of a line corresponding to a W^+ , W^- , or a Z^0 boson, and a set of rules for the calculation of probabilities can be formulated for these diagrams as well.

Interestingly, according to the standard model, the field theories for weak and electromagnetic interactions are parts of a broader, *unified* theory referred to as the theory of *electroweak* interactions. According to this broader view, the distinction between electromagnetic and weak interactions, which is apparent in interactions at relatively low energies, ceases to be a fundamental one at much higher energies.

While the strong interaction between hadrons may be explained, to some extent, with the mesons as the mediating particles, the hadrons themselves are bound configurations of quarks, where the quarks are the participants in strong and weak interaction processes involving hadrons at a basic level, and the question arises as to what mediates the interaction between the quarks. The theory postulates the existence of another kind of mediating particles, the *gluons* which are massless and chargeless particles of spin one like the photons. A gluon, however, carries colors with it so that color is conserved in the gluon-mediated quark interactions. The fact that the gluons, in spite of being massless particles, are not independently observed is because they remain confined within a small volume by an intrinsic feature of the mechanism of the interaction.

Finally, the gravitational interaction is mediated by another postulated elusive particle, the *graviton*. The theory requires the graviton to be a massless and chargeless particle of spin two. However, the fact that the gravitational interaction is the weakest of all the basic interactions makes it difficult for the graviton, the quantum of the gravitational field at the microscopic level, to be detected experimentally though *macroscopic* effects of

the field, at a classical level, are easily observed, and the search for *gravitational waves* has recently been rewarded with success. This contrasts with the electromagnetic field where classical states of the field are routinely observed while effects relating to non-classical, or quantum states are *also* observed with specially designed set-ups.

18.8.9.8 Symmetries and the conservation laws

The principle of conservation of linear momentum in mechanics is a consequence of the *homogeneity of space* where the latter means that the interaction between any two or more particles remains the same if these particles are all given the same displacement in space. This is expressed by saying that space is symmetric with respect to the operations of translation, and the interaction between particles is *invariant* under this symmetry operation.

More generally, any conservation law obeyed by any specified interaction is a consequence of one or more such symmetry principles where the energy function or the *Hamiltonian* characterizing the interaction (the energy function or Hamiltonian in quantum theory was introduced in sec. 16.5 in the context of the simple harmonic oscillator) is invariant under one or more symmetry operations.

For instance, the Hamiltonian describing the strong interaction between the hadrons (or, more fundamentally, the corresponding fields) is invariant under the operation of *space inversion*, i.e., the transformation from a right-handed to a left-handed co-ordinate system, and this results in the conservation of *parity* in the strong interactions. The weak interaction Hamiltonian, on the other hand, does not have this inversion symmetry, i.e. is not invariant under the operation of space inversion, as a result of which, parity conservation is violated in the weak interactions.

18.8.9.9 The Higgs field and the Higgs boson

The standard model is, in a sense, a mathematically incomplete one and needs to be complemented with an additional field, the *Higgs field*, along with the associated *Higgs boson*. The Higgs field gives consistency to the standard model by eliminating a number

of divergences in the probabilities of physical processes calculated with its help (a model with such unwanted divergences is termed a *non-renormalizable* one), thereby elevating it to the status of a *renormalizable* theory. In addition, it accounts for the *masses* of other elementary particles, thereby playing a significant role in the way the standard model may be made to describe the real world. One concept of central importance relating to the role of the Higgs field in the standard model is that of *spontaneous symmetry breaking* since it is precisely by means of such a symmetry breaking that the Higgs field causes different elementary particles to be endowed with different masses.

The particle associated with the Higgs field is the Higgs boson, being in the nature of a quantum state of the latter. Its production has been reported in high energy experiments.

18.9 The physics of molecules

The structure of all matter around us involves a hierarchy of bound states. Thus, an atom is a bound state of electrons with the nucleus, while the nucleus itself is a bound configuration of nucleons. At the next level down the hierarchy, nucleons are bound configurations of quarks. Moving *up* the hierarchy, on the other hand, one finds atoms bound together to form *molecules*. Finally, the molecules are held together, again in a bound configuration, in a solid. In a liquid or a gas, however, the molecules are at the very top of the hierarchy of bound configurations. In a liquid there arise only weak and transitory bindings between groups of molecules, each of which moves about while retaining its individual identity, while in a gas, even this transitory binding is absent.

In general, the binding energy per particle decreases as one moves up the hierarchy. Thus, the binding energy per nucleon in a nucleus is of the order of a few Mev, while the binding energy of an electron in an atom is only of the order of a few eV, being larger for heavier atoms as compared to the lighter ones. Molecules have, in general, a lower excitation energy and binding energy per particle compared to those of atoms. In other words, when a supply of energy breaks up the binding at the upper levels of the hierarchy the bound structures at the lower levels persist.

We will now have a look at how atoms are bound together to form molecules. A useful and convenient way of describing the bindings between atoms is to imagine that they are held together by means of *bonds* to form molecules. A bond, however, is not a physical object joining up the atoms - it is just a way to describe a configuration of the atomic nuclei and the electrons that has a lower energy compared to the atoms separated from one another by a large distance.

18.9.1 The binding of atoms in molecules: molecular bonds

18.9.1.1 The ionic bond

The simplest example of molecular bonding is the formation of the *ionic bond*. Consider, for instance, a sodium and a chlorine atom separated by a large distance and imagine the following processes to be made to occur one after another : (a) supplying energy to the sodium atom so as to split off an electron from it, thereby producing a Na^+ ion, (b) allowing the electron to get bound to the chlorine atom, forming a Cl^- ion, with the release of some energy: a chlorine ion is more stable than a chlorine atom and an electron separated from each other since the chlorine atom actually exerts an attractive force on the electron because of the incomplete screening of its nuclear charge by the extranuclear electrons; (c) allowing the positively charged sodium ion and the negatively charged chlorine ion to form a bound configuration, with the release of some energy, since the two exert an electrical attractive force on each other. At the end of these three steps, one has a bound configuration, the *NaCl molecule*, with some net energy (nearly 4.2 eV) released in the process, which is thus the *binding energy* of the NaCl molecule.

The term binding energy needs some explanation here. One may consider a NaCl molecule to be a bound configuration of a Na atom and a Cl atom, in which case the binding energy is 4.2 eV as stated above. Or one may alternatively consider the NaCl molecule as a bound configuration of an Na^+ ion and a Cl^- ion, in which case the binding energy turns out to be nearly 5.7 eV. In other words, one has to specify the *reference configuration* with respect to which the binding energy is defined.

One says that the NaCl molecule is bound by means of an *ionic bond*. The equilibrium

separation between the Na^+ and the Cl^- ions is referred to as the bond length. It corresponds to the minimum potential energy of the two ions when considered as a function of the distance between the two. At larger distances (r), the potential energy, increases to zero like the electrostatic potential between two opposite charges ($-\frac{e^2}{4\pi\epsilon_0 r}$) while at smaller distances it increases once again because the nuclei of the two ions repel each other and, additionally, the electron density clouds (see sec. 18.5.1) of the two ions tend to come too close together - something that the exclusion principle does not allow.

Fig. 18-14 depicts schematically the nature of variation of the potential energy of a Na^+ - Cl^- system relative to a Na-Cl system (with the Na and Cl atoms separated by a large distance) as a function of the separation between the Na^+ and Cl^- ions. The distance at which the potential energy is minimum corresponds to the equilibrium separation (≈ 0.24 nm) between the ions in a NaCl molecule, the latter being a bound configuration of the Na^+ and Cl^- ions held at this equilibrium separation.

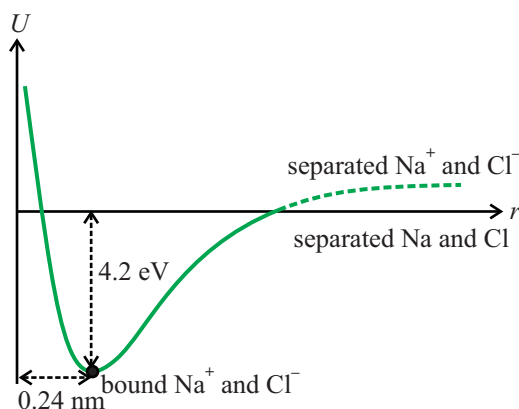


Figure 18-14: Energy (U) of a pair of Na^+ and Cl^- ions plotted as a function of their separation (r), relative to the energy of the Na and Cl atoms separated from one another; the minimum of the curve corresponds to the equilibrium configuration that can be described as the NaCl molecule; the zero of the energy scale is chosen to correspond to a Na and a Cl atom separated from one another, with respect to which the binding energy (i.e., the energy of the bound NaCl molecule is -4.2 eV); with reference to the separated Na^+ - Cl^- configuration, the binding energy is larger; the 'bond length' is 0.24 nm.

18.9.1.2 The covalent bond

Consider a hydrogen molecular ion (H_2^+), made up of two protons and one single electron. None of the two protons gets to enjoy the company of the electron exclusively, since the electron is equally shared between the two protons as in fig. 18-15, where a pair of classical orbits of the electron are shown schematically. In reality, the state of the electron is determined by the principles of quantum theory where the concept of an orbit is not a well defined one. It still possesses a measure of validity, though, since the variation in the probability of finding the electron at various different points can be represented by a smeared electron density cloud and the region where the cloud is most dense bears a resemblance to the electron orbit described classically. In this sense, the electron cloud representing the quantum state of the electron is termed an *orbital* (see sections 18.5.1 and 18.4.2 introducing the concepts of electron clouds and orbitals).

The orbital of the single electron in the hydrogen molecular ion (corresponding to the state of lowest energy for the system made up of the two nuclei and the single electron) provides for its binding by virtue of the attractive force it feels from the two nuclei, overcoming the repulsive interaction between the latter.

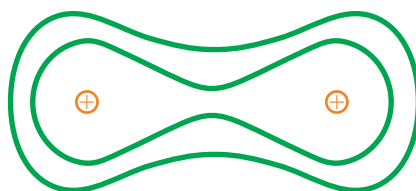


Figure 18-15: Depicting schematically two possible classical orbits of a shared electron in hydrogen molecular ion; quantum mechanically, the concept of an orbit is replaced with that of an orbital, where the orbital corresponding to the lowest energy of the system describes the hydrogen molecular ion; essentially the same orbital, only slightly modified, is shared by the second electron in the hydrogen molecule.

The hydrogen molecular ion, being a charged object, draws an electron towards itself whenever it finds one nearby, thereby leading to the formation of a hydrogen *molecule*, with *two* electrons shared by the two protons in the molecule. Each of the electrons forms an orbital similar to that of the single electron in the H_2^+ ion, which is modified

somewhat by the Coulomb repulsion between the electrons. In spite of this Coulomb repulsion between the electrons, however, the hydrogen molecule is a configuration of lower energy compared to a hydrogen molecular ion and an electron separated from each other. The fact that two electrons tend to form the same orbital, forces their spin magnetic quantum numbers (m_s) (see sec. 18.3.3) to be different in accordance with the Pauli exclusion principle, thereby implying that one of the two spins is *up*, while the other is *down* in the shared orbital.

Both the two electrons in the hydrogen molecule seek out the same orbital, i.e., the quantum state with the same spatial structure of the wave function. This wave function can be either a *symmetric* or an *antisymmetric* one with respect to the plane bisecting perpendicularly the line joining the two nuclei. Of these, the former corresponds to a lower value of energy since the probability distribution corresponding to this orbital favors to a larger extent those regions around the two nuclei where the potential energy of the electron assumes a lower value.

This configuration of two electrons in the same orbital with opposite spins, shared between two atoms (or, more precisely, ions) is termed a *covalent bond*, which is responsible for the binding of a large number of molecules.

The covalent bond may be a *non-polar* or a *polar* one. In a non-polar bond, the two electrons are shared symmetrically so that the making of the bond does not result in an electric dipole moment in the molecule. On the other hand, a polar bond is, to some extent, analogous to an ionic bond in that the electron density cloud representing the probability distribution of the two electrons forming the bond has an asymmetric structure, with an excess of charge concentrated near one of the atoms and a corresponding deficit near the other. The water molecule, for instance, possesses an electric dipole moment that accounts for the large dielectric constant of water and the ability of water to break up the bonds of other molecules.

When two or more atoms form a molecule by means of covalent bonds, each atom participating in a bond contributes one of the two electrons forming the bond from its

pool of valence electrons, i.e., those in the last unfilled subshell. The remaining electron forming the bond is similarly contributed by some other atom participating in the bond from its own pool of valence electrons.

Of the valence electrons in an atom, only those are capable of being shared in a covalent bond which are not themselves tied up in the atomic orbital in an up-down combination (such electron pairs in an up-down combination in the valence shell are commonly referred to as *lone pairs*). For instance, of the four valence electrons in the 2p subshell of oxygen, two are tied up in an up-down combination while the remaining two are free to participate in a covalent bond, which is why oxygen can form only two and not four such bonds.

The quantum theory of valence electrons, with some of these forming lone pairs and some others available for the formation of bonds, however, involves more elaborate considerations. In these paragraphs I present a simplified view of the formation of covalent bonds.

.

Since each element can contribute only a certain fixed number of electrons to form covalent bonds, these bonds are characterized by the *saturation* property, which means that any given atom cannot form bonds with more than a certain maximum number of other atoms.

Since the orbitals of the valence electrons in an atom have definite orientation properties in space, the covalent bonds have pronounced *directivity properties*, i.e., given the position of one of the atoms forming a bond, the line joining it to the other atom cannot be in any arbitrarily chosen direction. By contrast, given the position of the Na atom, for instance, forming an ionic bond with a Cl atom, the line joining the two can be in any direction whatsoever.

18.9.1.3 The hydrogen bond

The ionic bond and covalent bonds of various types account for the binding of an overwhelming majority of molecules. The typical bond energy (i.e., the energy required to break a bond) in each of these types of bonding is of the order of a few eV. Compared to these the *hydrogen bond* is a much weaker one, where the typical bond energy is only of the order of a fraction of an electron volt. The hydrogen bond is found in hydrogen-containing compounds, especially in long-chained organic molecules like the DNA and protein molecules. The hydrogen bond also plays a significant role in *intermolecular binding* in numerous liquids and solids.

Consider a configuration consisting of a hydrogen atom (H) and two other atoms (A and B) in its vicinity such that H is covalent-bonded to A, where A is a strongly electronegative atom. The covalent bond is then of a polar type, with the positive end of the dipole located near the hydrogen atom. This positive charge then induces a dipole moment in the atom B by drawing the electron charge-cloud of the latter toward itself. The resulting dipole-dipole interaction is responsible for the binding of the configuration $B \cdots H-A$, where \cdots represents the hydrogen bond. The formation of the bond requires that B should also preferably be an electronegative atom. With the bond-forming (unpaired) electrons in the valence shell of B engaged in forming bonds with some other atoms, it is a lone pair in B that gets affected by the positively charged hydrogen in forming the hydrogen bond.

The hydrogen bond occurs in protein molecules, being relevant in the *secondary structures* of the proteins. A protein molecule consists of a polypeptide chain (the *primary structure*), with oxygen atoms, attached to carbon atoms in the backbone of the chain, sticking out from the backbone, as also with hydrogen atoms, attached to nitrogen atoms in the backbone, similarly sticking out. The secondary structure of the protein involves a twisting and local folding of the chain into certain preferred configurations, where the structure gets support and stability from hydrogen bonds being established between the hydrogen atoms (covalently bonded to nitrogen atoms) and the oxygen atoms, with their lone electron pairs drawn out towards the hydrogen atoms having an

excess of positive charge in them (see fig. 18-16). The secondary structure of a protein molecule is relevant to a large extent in determining its *tertiary structure*, which consists of a global coiling and folding of the secondary structure in an aqueous medium, where this tertiary structure determines the functioning of the protein as a biological molecule.

The hydrogen bond is also of relevance in the *double helix* structure of the DNA molecule where it exists in the form of bridges between the two strands of the double helix. A similar role is played by the hydrogen bond in other long-chained polymers as well where it gives stability to the structure of the polymer chain.

Hydrogen bonding between water molecules in liquid water is responsible for some of its exceptional properties.

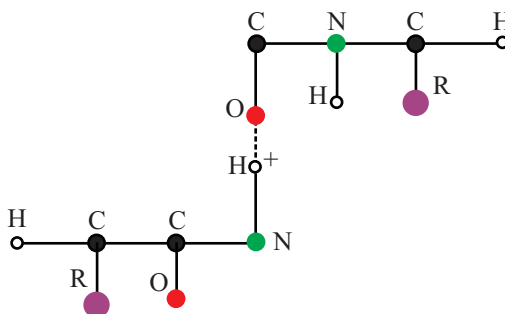


Figure 18-16: Hydrogen bonding in a protein molecule, only a small part of the molecule in its secondary structure is shown; bond lengths and bond angles are ignored.

18.9.2 Stationary states of molecules: molecular excitations

A molecule is made up of atoms, each of which, in turn, is made up of a nucleus and the electrons circulating in a number of orbitals. What is important to note in connection with the electronic orbitals is that, at least some of these get altered to a considerable extent as the atoms interact with one another to form molecules. For instance, the orbitals of the electrons that form the covalent bonds differ greatly from the orbitals of these same electrons in isolated atoms.

In order to describe the states of a molecule, one has to refer to these molecular or-

bitals. Unlike an atomic orbital, a molecular orbital cannot be described in terms of the quantum numbers n , l , m_l , and m_s . However, the broad picture remains unaltered: *stationary states* of the molecule corresponds to certain specific distributions of the electrons among the orbitals, each stationary state being characterized by a specific value of the energy of the molecule and by a set of other appropriately defined quantum numbers.

Among the stationary states, the one with the lowest value of the energy is the *ground state* of the molecule, which may be excited to states with higher energies. Conversely, a molecule in one of these excited states, may make a transition to a state with a lower energy. Such *molecular transitions* are accompanied by the absorption and emission of energy from the molecule. For instance, a photon may be emitted in a process of molecular de-excitation from a higher to a lower energy state.

However, the excitation and de-excitation of a molecule does not always involve a change in the occupancy of its electronic orbitals. Indeed, a comparatively large amount of energy is involved in the alteration of the electronic state of a molecule which is why such *electronic transitions* of molecules are relatively rare compared to changes involving *other* features of their states. These other feature are the ones relating to states of *vibrational* and *rotational* motions of the molecules.

The nuclei of a molecule can vibrate about their mean positions in certain characteristic *modes* with associated characteristic frequencies, where a typical mode behaves effectively as a harmonic oscillator (of frequency, say, ω), analogous to the standing wave modes of an electromagnetic field within an enclosure. In other words, the quantum mechanical description of a typical vibrational state is described by a vibrational *quantum number* (say, n_v , which is a non-negative integer), where the energy of the vibrational state is given by (refer to eq. (16-20))

$$E = \hbar\omega(n_v + \frac{1}{2}). \quad (18-30)$$

Such a description of a vibrational state is found to be relevant in the cataloging of

possible quantum states of a molecule and the possible transitions between these states. There can exist a number of different vibrational states for any given electronic state of the molecule and the energies associated with electronic transitions happen to be *large* compared with those for vibrational transitions. In other words, one can, from a practical point of view, make a neat separation between electronic and vibrational states in a complete description of the state of the molecule and of the transitions between these states.

The possibility of such a separation between electronic and vibrational states relates to the difference in typical *time scales* associated with changes in electronic and vibrational configurations of a molecule. Considering a diatomic molecule for the sake of simplicity, and adopting a classical approach for the time being, one can assume that, as the inter-atomic separation changes in the course of molecular vibrations, the electronic configuration, which can change over a much shorter time scale, adjusts itself instantaneously to the separation so that, for each and every value of the separation (r), one has some well defined value of the electronic energy (the ground state energy of the electrons in the quantum description), which leads to the definition of an effective potential energy function $V(r)$ of the molecule determined by the instantaneous electronic configuration depending parametrically on r . The separation (r_0 , say) for which this effective potential energy attains its minimum value then determines the vibrational equilibrium of the molecule, around which the vibrational motions can occur with various possible values of the energy. The latter includes the electronic energy (corresponding to separation r_0) as a constant component.

Similarly, the molecule as a whole can rotate like a rigid body about any given axis, and the typical energies associated with such rotational motion are small compared with the energies associated with the electronic orbitals and also, to a lesser extent, small compared to the energies of vibrational motion. The rotational state of the molecule is characterized by its *angular momentum*, where the energy of rotation in a state with a given angular momentum is characterized by a certain non-negative integral quantum number J , and several such rotational states, differing in their angular momenta and

energies, can exist for any given electronic and vibrational state.

In summary, the quantum mechanical stationary states of a molecule can be described, in a first approximation, in a simple scheme in which a state can be described in terms of three *independent* sets of parameters, one each for the electronic state, the vibrational state, and the rotational state. The energy of a stationary state can be expressed in the form

$$E = E^{(1)} + E^{(2)} + E^{(3)}, \quad (18-31)$$

where the three terms on the right hand side correspond to the electronic, vibrational, and rotational state of the molecule. Such a description is a meaningful one since, in general, one has $E^{(3)} \ll E^{(2)} \ll E^{(1)}$, and transitions are commonly observed involving changes in $E^{(3)}$ alone, for fixed values of $E^{(1)}$ and $E^{(2)}$. These are the rotational transitions of the molecule while transitions involving $E^{(2)}$ (vibrational transitions) and $E^{(1)}$ (electronic transitions) are also possible.

A large body of experimental observations relating to molecules and their spectra can be explained, in approximate terms, in this simple scheme, where the spectrum of a molecule consists of the frequencies of radiation emitted (emission spectrum) when the molecule makes transitions from higher to lower energy states, or of the frequencies of radiation that the molecule absorbs (absorption spectrum) in making transitions from lower to higher energy states. In the case of rotational and vibrational transitions, these frequencies usually lie in the infra-red and the microwave ranges of the electromagnetic spectrum.

However, observed spectra of molecules can be explained on the basis of the above scheme only if appropriate *selection rules* are taken into account. A selection rule is arrived at from the basic principles of quantum theory on the basis of which the *probabilities* of various transitions can be calculated. It is found from these rules that the probabilities of certain transitions are negligibly small compared to others. In other words, only certain transitions, in which the changes in the relevant quantum numbers

obey certain restrictions are *allowed* by the rules of quantum theory. These restrictions are termed the selection rules, and the transitions that conform to these selection rules are referred to as *allowed* ones.

In most cases, the observed transitions are caused by the interaction of the molecule with an electromagnetic field that may be present in the surroundings or, in the absence of an external field, with the *vacuum* field which is always present in the background. However, since the molecule is electrically neutral, the interaction with the electromagnetic field can occur only through the electrical dipole moment, ignoring the higher multipole moments of the molecular charge distribution. This means that, primarily, the transitions from one molecular state to another require the molecule to possess a dipole moment. Such is the case of a *heteronuclear* diatomic molecule where there is an asymmetrical charge distribution, with an attendant dipole moment. For such a molecule, there can occur a transition involving a change in the vibrational quantum number n_v by unity ($\Delta n_v = \pm 1$, the selection rule for vibrational transitions), with the electronic configuration remaining unchanged. Transitions of this type are observed in the infrared region of the electromagnetic spectrum. For a *homonuclear* diatomic molecule, on the other hand, there is no permanent dipole moment in virtue of a symmetrical charge distribution, and an electromagnetic interaction can take place only through the dipole moment *induced* by the electric field. Such an interaction causes a transition in the electronic configuration with, possibly, an attendant change in the vibrational state as well (satisfying the selection rule $\Delta n_v = \pm 1$; the case $\Delta n_v = 0$ corresponds to a pure electronic transition), observed in the optical or ultraviolet range.

In other words, vibrational transitions are of two types: those occurring without any change of the electronic configuration (example: heteronuclear diatomic molecules) and those accompanying electronic transitions (example: homonuclear diatomic molecules). A vibrational transition, on the other hand, is usually attended with a rotational transition involving a change in the rotational quantum number J , where the change in the rotational energy is generally small compared to that of the vibrational energy. Consequently, vibrational transitions are commonly observed to result in a *band spectrum*, where the band structure arises in virtue of the rotational transitions.

Deviations from the above simple scheme of description of molecular states and molecular spectra may occur because of the interdependence of the rotational and the vibrational states in certain situations, as a result of which, a description in terms of independent parameters relating to vibrational states on the one hand and rotational states on the other no longer remains valid.

18.10 From molecules to solids

While a molecule is a bound state of a few atoms or ions, a *solid* is likewise a bound configuration of a large number of ionic, atomic, or molecular units. A piece of sodium chloride crystal, for instance, is an object where sodium and chlorine ions are held together by ionic bonds. When molten sodium chloride or an aqueous solution of it is cooled, a sodium chloride crystal gets formed because of the fact that a geometrically ordered arrangement of sodium and chlorine ions fastened to one another with ionic bonds is a configuration of lower energy (or, more precisely, lower *free* energy) than other possible configurations, which means that such an arrangement is more durable than other possible ones, requiring a greater amount of energy to break it up.

Suppose a solid is formed from N number of atomic or molecular constituents (in the case of ionic constituents, corresponding neutral atoms or atomic groups are considered, like the NaCl molecule made up the sodium and chlorine ions in sodium chloride crystal) held together by means of bonds such as covalent ones where any one constituent is usually bonded to only a few neighboring ones because of the finite range of the bonding (ionic bonds, however, are relatively long range ones). The energy necessary (say, per mol of the material) to take apart all the constituents to a large distance from one another, with each individual constituent remaining at rest, is termed the *cohesive energy* of the solid.

The mechanisms responsible for cohesion in solids may be more varied than those involved in molecular bonding. For instance, the interactions of free electrons in a conducting material with the ionic cores forming the lattice structure have a dominant role in the cohesion of the material. This is analogous to the covalent bonds in molecules,

with the important difference that the electrons involved in the bonding (referred to as the *metallic* bonding) are *delocalized*, their wave functions being extended throughout the crystalline structure, while the electrons in a covalent bond are *localized* (they are however, delocalized with reference to the constituent atoms). In this sense, a crystalline conductor can be described as a *giant molecule* with covalent-like bonding. Covalent bonding with localized electrons are also relevant in a large number of solids. Bonding mechanisms other than the covalent, ionic, and the metallic bondings, such as the hydrogen bond and the *Van der waals interaction* (interaction between fluctuating electrical dipoles) are also found to be operative in the cohesion of solids.

In contrast to a solid, a gas is characterized by *zero binding*, which means that no energy is to be supplied to a gas to enable its molecules to disperse to infinite distances from one another, as the molecules are released from a closed container. Though there does exist weak interactions between the gas molecules, the thermal kinetic energies of these molecules are sufficient to overcome the effect of these interactions.

A *liquid*, on the other hand, is of an intermediate nature in that there arise *temporarily bound local groupings* among the molecules of the liquid. In other words, a liquid molecule finds itself bound, from time to time, with a number of other molecules (but not to all the liquid molecules at a time, as in a solid) in a temporarily formed conglomerate where the latter is transitory in nature and breaks up after a time, when the molecule under consideration gets attached to a conglomerate formed afresh. Once again, this involves a bonding between the molecules, but such bonding is relatively weak compared to that in a solid. An example of bonding in a liquid is provided by the *hydrogen bonding* among water molecules. This hydrogen bonding is of great relevance in being responsible for a large number of remarkable and exceptional properties of water as a liquid.

Chapter 19

Electronics

19.1 Introduction

An *electronic* circuit differs from an *electrical* one in that it makes use of *non-ohmic* elements in an essential way. Ordinary resistors, inductors, and capacitors are *ohmic* (or 'passive') elements for which the voltage across an element is proportional to the current through it where, for AC voltages and currents, the relation of proportionality holds between the respective root mean squared (RMS) values. The ratio between the voltage and current, known as the resistance for an ordinary resistor (say, a piece of wire) and impedance for an inductor or capacitor (see chapter 13) is a constant, independent of the voltage or current. By contrast, the relation of proportionality does not hold for a non-ohmic (or 'active') element like, for instance, a *p-n junction diode*.

Other examples of non-ohmic elements are the *light emitting diode* (LED), the *bipolar junction transistor* (BJT; or the *transistor* as it is commonly known), the *field effect transistor* (FET), and the *silicon-controlled rectifier* (SCR), while a host of other devices are also in wide use in electronic circuits of a vast range in variety and complexity.

Of course, it is difficult to find a circuit of practical value that is purely electrical or purely electronic according to the above point of view. The term electrical or electronic, when used to describe a circuit, tells us whether the functioning of the circuit is depen-

dent, in the main, on ohmic or on non-ohmic elements.

A universal feature of the devices that play a crucial role in electronic circuits is that they are made up of *semiconducting* materials like pure or doped germanium (Ge), doped silicon (Si) (it is difficult to produce a silicon crystal of adequate purity), or gallium arsenide (GaAs). Devices based on such materials can be made to exhibit a fascinatingly rich electrical and electro-optic behavior through variations in their fabrication techniques, engendering actual and potential applications over an almost *unlimited* range.

Modern day electronic devices are almost universally made up of *integrated circuits*, where an integrated circuit is a piece of miniature size, including active and passive elements connected to form some specific circuit design, the connections being made with the help of miniature conducting paths laid out on the same small piece of *substrate* on which the other circuit components are formed by means of surface deposit and etching. An integrated circuit (IC) is commonly referred to as a ‘chip’ where chips can be of various sizes made up of only a few circuit components or of a staggeringly large collection of those.

In this brief introduction to semiconductors and electronic circuits I will first outline a number of basic concepts relating to the electrical properties of semiconductors and will then touch upon the principle of working of the p-n junction diode and the transistor. After a short introduction to operational amplifiers and oscillators, I will finally outline a few basic principles of *digital* electronics.

Electronic circuits can be classified into *analog* and *digital* ones. Once again, this classification is nothing more than a way of looking at things since only few circuits of importance can be labeled as being purely analog or purely digital. The terms analog and digital apply more appropriately to *parts* of a circuit rather than to a circuit as a whole. Simple examples of analog circuits are the rectifier and the voltage amplifier, which I will introduce in sections 19.3.6 and 19.4.11. In later sections of this chapter I will explain what distinguishes a digital circuit (or part of a circuit) from an analog one. The former (see sec. 19.7) are made up of simple building blocks called *logic gates*

and *flip-flops* (the latter being based on the former). A number of these gates and flip-flops will be introduced in this chapter. Integrated digital circuits are now ubiquitous in devices ranging from large computers to ordinary home appliances.

Electronics is an applied subject of vast proportions, where the scenario changes almost every day. Since an acquaintance with the principles of operation of electronic circuits forms an integral part of any course in basic physics, I include in this chapter a sketchy outline of a few basic electronic circuit components and devices, hoping that this will serve to launch you into the fascinating world of electronics.

19.2 Electrical properties of semiconductors

A *semiconductor* is a crystalline material whose electrical properties are, in a sense, intermediate between those of a *conductor* and an *insulator*. To start with, a few words are necessary to explain the relevance of *energy bands of electrons* in a crystalline material in determining its electrical properties.

A crystalline material is made up of atoms arranged in a *regular* or periodic three dimensional structure. In such a regular structure, the electrons, instead of being attached to or localized around individual atomic nuclei, can *hop* around from one atom to another. This contrasts with an electron in an isolated atom where the electron is captured in a *bound* orbital around the nucleus (see chapter 18). In a molecule made up of more than one atoms, on the other hand, an electron can be *shared* between the atoms where its orbital covers all the atoms in the molecule. The electrons in a crystal behave in a somewhat similar fashion.

19.2.1 Energy bands of electrons in a crystal

As explained in chapter 18, a neutral atom is made up of the nucleus together with a number of electrons distributed in shells and subshells, the electrons in outer shells being loosely bound to the nucleus compared to inner ones. When a large number of such atoms are brought together to form a crystalline structure, the electrons get

detached from the individual atoms so as to form a *common pool* of electrons, each member of the pool being *shared* by all the atoms in the crystal. This is made possible by the fact that the atoms are arranged in the crystalline structure with a periodic regularity whereby the rules of quantum theory require the orbitals of the electrons to embrace all the atoms constituting the crystalline structure.

So, strictly speaking, one should talk of *ions* instead of atoms making up the skeleton of the crystal since the atoms donate their electrons to the common pool. Each electron in this common pool experiences the electrostatic pulls from these ions and also interacts with all the other electrons moving around. In an approximate sense, one can ignore this mutual interaction among the electrons and still arrive at meaningful results relating to the electrical properties of the crystalline substance under consideration. In other words, each electron moves around in the lattice independently of the other electrons to a good degree of approximation. The rules of quantum theory can be invoked to work out the stationary states of such an electron and the corresponding energy values.

As in the case of an electron in an isolated atom, the possible energy values of an electron in a crystalline structure can be depicted pictorially with the help of a number of horizontal bars, each bar representing an energy level of the electron. One general result to come out from quantum theoretic considerations is that the energy levels of an electron in a crystalline medium form a number of *energy bands*, with *forbidden energy gaps* separating the successive bands.

This is depicted schematically in figure 19-1, where a number of bands have been shown, with the energy increasing from bottom upward. Each band is made up of a large number of energy levels, represented by the horizontal bars.

Of the large number of discrete energy levels packed close together so as to form any of the bands, only a few are shown in the figure. The successive levels within a band are so close to one another that the corresponding energy values can be assumed to vary continuously from the bottom edge of the band to the top edge. This is a consequence of the fact that the volume of the crystal is almost infinitely large compared to a volume

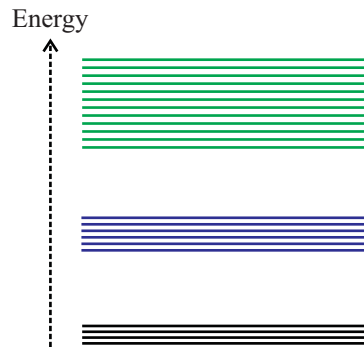


Figure 19-1: Schematic representation of electronic energy bands in a crystal; each band is made up of a large number of energy levels; successive bands are separated by energy gaps; the widths of bands increase with energy.

of atomic dimensions.

One more fact of relevance is that the width of the bands increases along the energy scale. Thus, the lowest bands are thin, with comparatively large gaps in between, while the ones higher up are wider, being comparable with the intervening energy gaps. In some instances, two successive bands may even become thick enough to *overlap* with each other, forming effectively a single energy band.

Looking at the single electron energy levels of an isolated atom and at the energy bands in a crystal made up of atoms of the same kind, one can discern a certain correspondence between these. Thus, in a certain sense, each band may be thought of as being produced by the ‘splitting’ of an energy level of an isolated atom into a large number of closely packed levels forming the band, where the number of levels in the band equals the number of atoms in the crystalline structure. The lowest band corresponds to the K-shell of the isolated atom and the next higher band to the L-shell, made up of the 2s and 2p orbitals, the latter being almost degenerate in energy so as to be represented by a single energy level for the isolated atom. The next two bands, correspond to the 3s and 3p orbitals of the isolated atom which differ only slightly in energy, as a result of which these two bands may overlap (the 3d band lies higher up in energy, though it may also overlap with a neighboring one). The actual widths of the bands and the energy gaps between them vary considerably from one crystalline material to another and there is

considerable variation in the overlapping of bands as well, figure 19-1 being only an illustrative example.

19.2.2 The filling up of the bands

A common feature of the possible orbitals of an electron in a crystalline environment is that each of the orbitals embraces all the atoms (or, more correctly, ions) in the crystal, i.e., is not localized exclusively around one single ion or a group of ions (by contrast, the orbitals are concentrated near isolated clusters of ions or atoms in an *amorphous* material). Apart from this common feature, however, the orbitals differ widely in their structure and a number of related features. Of particular relevance in this context are the so-called *valence band* and *conduction band*. But before talking of these two, I have to tell you how all the large number of electrons in a crystalline material fill up the successive bands.

All the electrons detached from the atoms making up a crystalline structure count up to a very large number. In an approximate sense, all these electrons move about through the material independently of one another, each in an orbital of its own, with an energy belonging to the various energy levels making up the energy bands. The question then comes up as to how the electrons are distributed among the energy levels.

In general, each energy level corresponds to a single orbital where the term orbital denotes a stationary state without reference to its internal or spin state. Since there can be two distinct spin states of an electron (with $m_s = \pm\frac{1}{2}$, see chapter 18), there can be two distinct stationary states for each of the energy levels. Recalling, then, that the electrons obey the Pauli Exclusion principle whereby no two electrons can be in the same stationary state, one concludes that there can be at most two electrons occupying each energy level (corresponding to an orbital) in the crystal.

The electrons fill up the energy levels from bottom upwards. If, by any chance, an electron finds itself in a level higher up in the energy scale with a lower level vacant, it loses no time in dropping down to the vacant level, giving up its energy either in the form of photons or as vibrational energy of the ions in the crystal. So, the levels are occupied

in a manner similar to how the single electron energy levels in an isolated atom are filled up: *bottom upwards, two in each level*.

What I outline here is a rather simplistic theory of electrons in a crystalline solid, where the interactions between the electrons is not taken into account. More precisely, the theory can be so formulated as to include the inter-electron interactions in an approximate manner in the form an average potential felt by each electron over and above the potential due to the periodic array of ions in the crystalline lattice, this being analogous to the screening effect felt by electrons in an atom. As we saw in atomic theory in chapter 18, a realistic account of the atom has to include the effects of electron-electron interaction beyond what the screening effect describes. Likewise, the interaction among the electrons in a solid assumes relevance in numerous situations of interest, though it is no less significant that the independent-electron approach (with the electron-electron interaction accounted for in terms of an average effective potential felt by each electron) works rather well, at least for a qualitative understanding of a large class of solid state phenomena. The simplified approach will be sufficient for our present purpose.

The above simple rule of level occupancy applies, strictly speaking, only when the temperature of the entire system is held at absolute zero. At a higher temperature, say T , some of the electrons may be *thermally excited* from lower to higher levels. On the average, the thermal energy that an electron can acquire from its surrounding electrons and ions is of the order of $k_B T$, where k_B stands for the Boltzmann constant. For the time being, however, we will stick to the simple rule of level occupancy stated above, looking at finite temperature effects later.

19.2.3 The valence- and the conduction bands

As all the electrons are accommodated in the energy levels in accordance with the above rule, some of the bands lower down in the energy scale are likely to have all their energy levels completely filled up. The highest band with all the levels occupied to the full is termed the *valence band*.

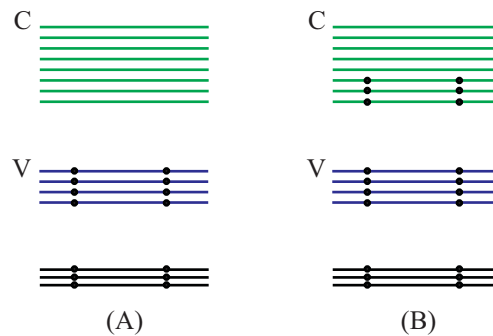


Figure 19-2: Energy bands in a crystal (schematic); in (A) the conduction band is empty, while in (B), it is partially filled, with two electrons filling up each level; in both the cases, the valence band (V) and the bands lower down in energy (only one is shown in each) are filled up.

The next higher band, which may be either completely empty or partly filled up, is referred to as the *conduction band*. Figure 19-2 depicts schematically the two alternative situations where, in (A), the conduction band C is empty while, in (B), it is partly filled up. In both the figures 19-2 (A) and (B), the valence band V and the bands lower down in the energy scale (only one such band being shown in each case) are completely filled up.

Electrons in the conduction band behave differently as compared to those in the valence band even as both are shared by all the atoms in the crystal structure. Imagine a weak electric field being set up in the crystalline medium. One expects an electron to be accelerated by this field and gain in energy. However, since all the energy levels in the valence band are filled up, an electron in the valence band *cannot* take up energy from the electric field without violating Pauli's principle - it has no vacant level to move to with this extra energy! The conduction band, on the other hand, has vacant levels in it and an electron lodged in it can take up energy from the electric field, moving to a higher level with this acquired energy.

Had the electric field been *strong* enough, even an electron in the valence band could have acquired sufficient energy to cross the energy gap between the valence band and the conduction band, thereby jumping up to the latter. However, such strong fields are not under consideration here since they are set up only under specially designed situa-

tions. Our purpose here is to see whether the crystalline material under consideration can respond to a *weaker* field by way of generating an electric current.

Electrons in the conduction band can generate such a response by a process of alternate acceleration and deceleration, thereby acquiring an average *drift velocity* due to the electric field. While the terms acceleration and deceleration refer to a classical picture of the motion of the electrons, they are quite convenient in the present context. The acceleration corresponds to the electron taking up energy from the field and moving over to a vacant higher level. The deceleration phase then follows as the electron quickly gives up its acquired energy to the ions in the crystalline medium, thereby enhancing the vibrational motion of the ions. The resulting drift velocity, when acquired by a number of electrons in the conduction band, leads to the generation of a *current* since the electrons are all charged particles.

Strictly speaking, the behavior of electrons in the crystalline medium, including their response to external fields, is to be described in quantum theoretic terms. However, it is convenient, and at the same time meaningful, to use a number of classical concepts in such a description. In such a mixed description, one has to be careful of not losing sight of the quantum features that are essential in the sense that these cannot be explained in classical terms.

Talking of the quantum features, it is useful to compare the motion of an electron in a crystalline environment with that of a free electron. The electron in a crystal lattice experience a *periodic* field of force (corresponding to a periodic potential) due to the ions arranged regularly in the crystal. Even so, its motion resembles that of a free electron in that the electron can move about throughout the volume of the crystal without being tied to a single ion or group of ions. Recall that the wave function of a free electron is in the nature of a plane wave with a certain wave vector. The wave function of the electron in the crystal differs somewhat from a plane wave in that the amplitude of the wave, instead of being a constant, gets modulated periodically in the crystal. What is more, the response of the electron to an impressed weak electric field can be described in terms analogous to ones relating to the response of a free

electron, with the difference that the effective *mass* (say, m^*) of the electron in the crystal depends on the wave vector (or, equivalently, on its energy), where the effective mass can vary from a small fraction of the free electron mass to several times the same.

In other words, electrons in the conduction band can be looked upon as *carriers* - negatively charged ones - in the generation of an electric current. Electrons in the valence band, however, cannot act as carriers since they cannot take up energy from the electric field. Looked at this way, the electron orbitals in the valence band are all 'frozen in' as compared with those in the conduction band which are, by contrast, 'free' - while the former cannot be changed or deformed by a weak electric field, the latter can easily be.

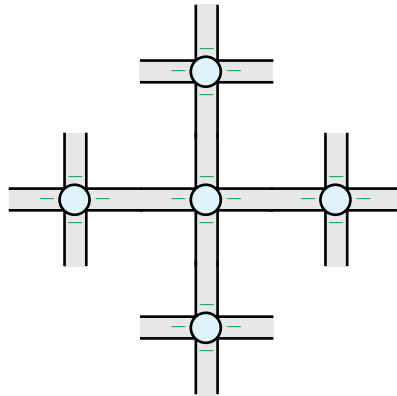


Figure 19-3: Bonds between neighboring atoms (more precisely, *ions*, represented by circles) in a crystal lattice; in germanium and silicon, each atom is bonded to four neighbouring atoms in a tetrahedral structure (not shown in this figure); the bonds are schematically represented by pairs of lines, each bond being formed of two electrons ('-' symbols); bonds linking one central atom with four neighbours are shown; the bonds are not physical structures linking the atoms, and electrons move from one bond to another, since these are shared by all the atoms forming the lattice.

The orbitals of the electrons in the valence band can, in a sense, be looked upon as forming *bonds* between neighboring atoms in the crystal structure (see figure 19-3), much like the covalent bonds in molecules. Each bond is made up of a pair of electrons coming from the common pool. The filling up of the valence band corresponds to all the

possible bonds having been formed. If, by any chance, an electron is taken out of the valence band then the resulting vacancy may be interpreted as one of the bonds lacking in one of the pair of electrons necessary for its formation. However, an electron from a neighboring bond can move in to fill up the vacancy (recall that the electrons all belong to the common pool), which means that the *vacancy* has now moved from one bond to a neighboring one.

Such a vacancy in a sea of electrons is termed a *hole*. The motion of a hole is similar to that of a positive charge and is in reality due to the successive motion of electrons in the opposite direction from bond to bond. The holes can acquire a drift velocity caused by an applied weak electric field. In terms of energy levels, a vacant energy level in the valence band allows an electron to fill it up by acquiring energy from the field and then releasing this energy to the vibrating ions in the crystal lattice. A number of such vacancies in the valence band thus generates a current that can be interpreted as being due to the motion of the holes - *positively* charged carriers.

The concept of holes is, strictly speaking, once again a quantum theoretic one. If one imagines, for instance, a filled band with just one electron missing in it, and considers from the quantum point of view the motion of the electrons in a weak electric field imposed on these, one ends up with a current that can effectively be interpreted as that due to a positive carrier with a certain effective mass (m^*), where the latter depends on the energy of the level in which the vacancy has been produced.

19.2.4 Electrochemical potential and the Fermi level

One other concept I want to talk of is that of the *electrochemical potential*. The *electrical* potential of a charged particle was introduced in chapter 11. The variation of electrical potential from point to point in a crystalline medium will give you the direction in which an electron, left to itself, will move in it. The electrical potential determines the electric field intensity, which acts as the motive force for the motion of the electron because the latter is a charged particle. Another reason why electrons will move in a medium is the variation in the *number density* of the electrons from point to point. This is the process

of *diffusion* which can occur regardless of the charge of the particle.

The diffusion current depends not only on the variation of number density but on temperature as well, and is effectively determined by a parameter termed the *chemical potential* - a quantity with the dimension of energy. More generally, if there exists an electric field together with a variation in the chemical potential then the resultant effect of diffusion and drift (i.e., the motion driven by the electric field) is described by a combination of the electrical and chemical potentials - the *electrochemical potential*. If, for instance, the electrochemical potential is uniform everywhere in a medium then there can be no net current in it, i.e., in other words, the diffusion current and the drift current (another name for the current driven by the electric field) must cancel each other.

In describing the distribution of electrons among the energy levels making up the bands in a crystalline material, I mentioned that the temperature could be assumed to be zero. This is a good enough assumption for typical crystalline materials since, in reality, the distribution of electrons at temperatures even as high as, say, 300 K does not differ appreciably from that at 0 K. With this in mind, we can approximate the electrochemical potential by its value at 0 K. This is referred to as the *Fermi energy* or as the *Fermi level*. The latter is depicted as a horizontal dotted or solid line placed appropriately among the energy bands. No energy level in a band higher than the Fermi level can be occupied at 0 K. Thus, in figure 19-2 (A) where the conduction band is empty, the Fermi level lies in between the valence and conduction bands while, in (B), it lies somewhere inside the conduction band.

In the literature one often finds the term Fermi level being used to denote the electrochemical potential itself at any given temperature T , rather than the limiting value of the latter for $T \rightarrow 0$. In the following, I will also indulge in this rather loose usage of the term, which will thus imply that the Fermi level will be a temperature-dependent parameter. For typical semiconductor materials, this dependence is a weak one.

19.2.5 Energy bands: summary

In summary, then, the electrons in a crystalline material form a common pool, donated by the atoms forming the crystalline structure, where the atoms reside in the crystal in the form of ions. Each electron moves about in the crystal structure more or less independently of the others, and can be in one of a large number of energy levels, which are organized into bands, with forbidden gaps in between. Each energy level can accommodate a maximum of two electrons with oppositely directed spins. The highest fully occupied band is termed the valence band while the next higher band, which may be either empty of electrons or partly filled, is termed the conduction band. A conductor is distinguished from an insulator or a semiconductor (see below) by having a partially filled conduction band at 0 K.

The electrons in the valence band, while belonging to the common pool, are, in a manner of speaking, frozen in, since these cannot take up energy from a weak electric field so as to form a current. An electron in the conduction band, on the other hand, is free to take up energy from a weak applied field and to participate in the setting up of a current.

Though the valence band is completely filled up at 0 K, at higher temperatures, a small number of the valence band electrons may move up to the conduction band, thereby generating *holes*, which act as positive carriers, thus contributing to the current that may be set up in the material. As we will see below, holes can also be formed by appropriate *doping* of the crystalline material, where electrons from the valence band are captured by dopant atoms.

A current can be set up in the crystalline material due to the action of an electric field as also due to a concentration gradient of the carriers. The resultant effect of the two is expressed in terms of the electrochemical potential or, ignoring the weak temperature dependence of the latter, by the *Fermi level*. The Fermi level of the electrons in a conductor, for instance, lies within the conduction band.

19.2.6 Conductors, insulators, and intrinsic semiconductors

With this background on energy bands in crystalline solids one can now attempt at distinguishing between conductors, crystalline insulators and *intrinsic* semiconductors. While metals are typical examples of conductors, insulators may be either amorphous or crystalline, an example of the latter being crystalline silicon dioxide. Examples of intrinsic semiconductors are pure crystals of germanium (Ge) and silicon (Si), both being tetravalent elements belonging to group IV of the periodic table. Later in the present chapter (sec. 19.2.7) I will introduce the *doped* semiconductors - materials used widely in electronic devices.

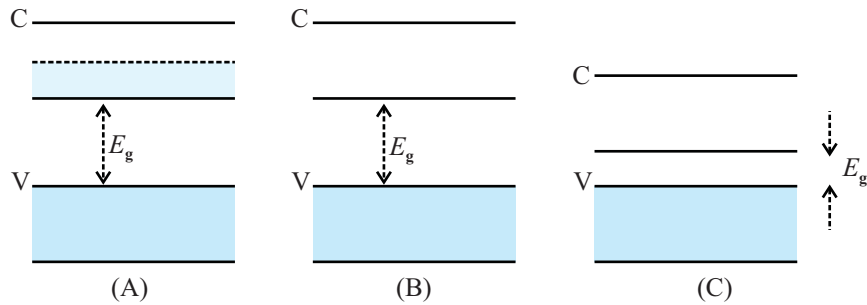


Figure 19-4: Schematic representation of band structures in (A) a conductor, (B) an insulator, and (C) an intrinsic semiconductor; shading indicates filled-up levels in a band; in each case only the valence band (V) and the conduction band (C) are shown; in (A) the conduction band is partially filled, while in (B) and (C), it is empty at 0 K; E_g denotes the band gap.

Figure 19-4 depicts schematically the energy band structures of a conductor, a crystalline insulator, and an intrinsic semiconductor. The band structure in (A), where the conduction band is partly filled up, corresponds to a conductor because the electrons in the conduction band act as negatively charged carriers, generating an electrical current when a weak electric field is set up in the material and, correspondingly, implying an appreciable *conductivity*. By contrast, (B) and (C), in both of which the conduction band is empty, correspond to an insulator and an intrinsic semiconductor respectively. The two are distinguished by the fact that, while in (B) the energy gap between the conduction band and the valence band is large compared to $k_B T$ at ordinary temperatures (say, ~ 300 K) it is much smaller in (C), being of the order of several times $k_B T$.

At 0 K, then, no carriers are available for either of the situations depicted in figure 19-4 (B), (C). At higher temperatures, on the other hand (the typical temperature I will refer to will be $T \sim 300$ K), the two situations imply distinct behaviors. The energy gap between the valence and conduction bands being large compared to $k_B T$ in figure 19-4 (B), very few electrons will be thermally excited to the conduction band at temperature T , and thus the material under consideration will behave as an *insulator*. Indeed, the temperatures required for enough number of electrons to be thermally excited so as to lead to appreciable conductivity far exceed the *melting points* of typical insulators.

On the other hand, for a material with an energy band structure as in fig. 19-4 (C), a number of electrons get thermally excited at higher temperatures (T) from the valence band to the conduction band owing to the fact that the band gap to be overcome is much smaller. While this number is still small compared to the number of available carriers in the case of a conductor, it is nevertheless sufficient for the material to have a small but appreciable conductivity. Recalling that, for every electron excited to the conduction band from the valence band, a hole is simultaneously generated in the valence band, one concludes that the conductivity will be a sum of contributions from the electrons in the conduction band *and* holes in the valence band. The directions of drift velocity generated by an electric field will be opposite for the two types of carriers, but the resulting currents will add up owing to the opposite signs of charges.

In summary, then, the band structures for conductors, crystalline insulators, and intrinsic semiconductors imply a considerably high value of the electrical conductivity for a conductor, negligible conductivity for an insulator, and a small but detectable conductivity for an intrinsic semiconductor.

Silicon and germanium are familiar examples of intrinsic semiconductors. Of these two, Ge crystals of sufficient purity can be obtained so as to show intrinsic conductivity at a temperature of the order of 300 K. Si, on the other hand, cannot be crystallized to a sufficient degree of purity so as to show intrinsic conductivity at such temperatures. Instead, the conductivity exhibited by a Si crystal arises due to impurities that are unavoidably present in it (see sec. 19.2.7 for an introduction to conductivity produced

by impurity centers in a crystal of an intrinsic semiconductor).

The distinction between a conductor and an intrinsic semiconductor lies not only in the magnitude of the electrical conductivity, but in the *temperature dependence* of the conductivity as well. For a conductor, the number of carriers is typically so large even at 0 K, that the fractional change in the number of carriers due to thermal excitation from the valence band is negligible. While this number remains almost unchanged, the *mobility* of the electrons (drift velocity acquired per unit magnitude of the impressed electric field) decreases with increasing temperature. As the temperature is made to increase, a comparatively larger number of electrons get *scattered* from the vibrating ions in the crystal lattice whose vibrations grow more and more vigorous with increasing temperature. This process of scattering results in an effective friction-like force impeding the motion of the electrons, causing the mobility, and hence the conductivity, to decrease at higher temperatures.

For an intrinsic semiconductor, on the other hand, this decrease in mobility is *overshadowed* by the increase in the number of carriers (electrons *and* holes) due to thermal excitation across the small band-gap. This results in an *increase* in the electrical conductivity of the material with a rise in temperature.

Incidentally, and interestingly, higher values of electrical conductivity imply higher values of *thermal* conductivity as well. Both are due to *transport processes* involving the same kind of carriers.

19.2.7 Doped semiconductors

Intrinsic semiconductors are, in reality, of little use as electronic devices. Crystals of a high degree of purity are difficult to produce and, more importantly, the conductivity of an intrinsic semiconductor, which is rather too low, depends only on the temperature. Hence it cannot be made to vary by convenient means according to necessity. Indeed, the rather pronounced temperature dependence of the conductivity is somewhat of a disadvantage since fluctuations in the ambient temperature cause an undesirable vari-

ation in the electrical behavior of the material.

A *doped* or extrinsic semiconductor is one in which a small amount of impurity of an appropriate kind is introduced in the crystalline material. The dopant can be either a group III element (e.g., aluminium (Al), gallium (Ga), indium (In)), termed an *acceptor*, or a group V element (e.g., phosphorus (P), arsenic (As), antimony (Sb)), termed a *donor*. The addition of these impurities in even very small concentrations leads to a dramatic increase ($\sim 10^8$ times) in carrier concentration (number of carriers per unit volume) and hence of electrical conductivity of a semiconducting material. What is more, a desired degree of increase in the conductivity according to necessity can be achieved by controlling the dopant concentration. Controlled doping makes possible a great flexibility in the operational features of semiconductor devices.

When a small concentration of a donor material is added to a germanium or a silicon crystal, the donor atoms take up positions in the crystalline structure similar to those of the host (Ge or Si) atoms. Each donor or host atom may be thought of as being located at the center of a regular tetrahedron (schematically shown in fig. 19-3 in a planar representation), with bonds extending toward the four corners of the tetrahedron which, in turn, are occupied by other atoms. Each atom contributes four electrons to the common pool of electrons forming the bonds, all these electrons filling up the valence band. For the host atoms this accounts for all the electrons in the two outermost orbitals (s^2p^2). For a donor atom, on the other hand, there are *five* electrons in the two outermost orbitals (s^2p^3), from which four go to the common pool, while the fifth electron remains loosely bound to the donor atom in a hydrogen-like orbital.

Just as the electron in a hydrogen atom can get knocked off on receiving a supply of energy from its surroundings, thereby ionizing the atom, similarly the loosely bound fifth electron of the donor atom can also get knocked off, thereby gaining the status of a free electron in the conduction band of the crystal. However, the difference between the two cases lies in the amount of energy necessary to detach the electron from the bound configuration - the energy necessary in the case of the donor atom being much smaller ($\sim 10^{-3}$ times) than that for the hydrogen atom.

At zero temperature this fifth electron fails to receive energy from its surroundings and remains loosely bound with the donor atom, but at higher temperatures (as low as ~ 40 K) almost all these electrons from the donor atoms implanted in the crystal structure get detached and form a pool of free electrons in the conduction band. The donor atoms are now singly ionized with a positive charge and the doped semiconductor acquires a conductivity owing to the pool of free electrons which can act as carriers. Evidently, the carrier concentration and hence the conductivity can be varied according to necessity by varying the dopant concentration.

Similar considerations, though differing in details, apply to doping by acceptor atoms. Once again, the acceptor atoms are located similarly to the host atoms in the crystal structure but now the two outermost orbitals contain *three* instead of four electrons (s^2p^1) and hence one more electron is needed to saturate all the four bonds around the acceptor atom. The shortfall can be made up by an electron moving in from a neighboring bond whereby a hole is created in the latter. At or near 0 K, this hole remains loosely bound to the acceptor ion (which now acts as an attracting center with *negative* charge because of the extra electron coming up from the neighboring bond) in a hydrogen-like orbital. However, even at a slightly higher temperature, the hole gets detached, becoming thereby a 'free' hole, like one in the valence band of an intrinsic semiconductor. In the process, it leaves behind a negatively charged acceptor ion. One can express all this by saying that an electron has been lifted from the valence band to an 'acceptor level' (see below), the latter representing the energy in the ionized configuration of the acceptor atom. As almost all the acceptor atoms get ionized, the holes generated in the valence band act as carriers, generating a conductivity for the doped crystal.

This, then, is how a doped semiconductor acquires an electrical conductivity. The materials resulting from the two kinds of doping are termed an *n*-type and a *p*-type semiconductor respectively corresponding to the respective signs of charge (negative and positive) of the carriers.

The energy diagrams for these two are shown in figure 19-5 (A) and (B). For the *n*-type

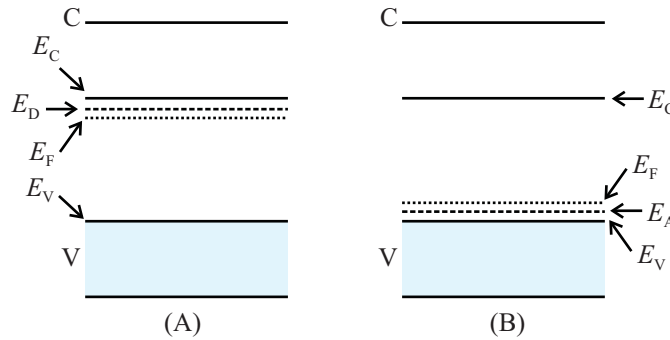


Figure 19-5: Energy band structure in (A) n-type and (B) p-type semiconductors; donor and acceptor levels (resp. E_D and E_A) are shown with dotted lines; the position of the Fermi level (E_F) for the intrinsic semiconductor is also shown in each case; the temperature is assumed to be in the range ~ 40 K to ~ 500 K; E_C and E_V denote the bottom of the conduction band and the top of the valence band respectively.

semiconductor the ground state energy of the loosely bound configuration involving the fifth electron of the donor atom is slightly lower than the energy corresponding to the bottom of the conduction band, which is why the electron can be easily detached and added to the conduction band (the necessary energy is only \sim several meV, where $1 \text{ meV} = 1.6 \times 10^{-22} \text{ J}$). This is shown in fig. 19-5(A) as an energy level referred to as the *donor level*. The physics of an n-type semiconductor is summarized by saying that the donor orbitals are occupied at $T = 0$ K, while almost all of them give up their electrons to the conduction band at even slightly higher temperatures.

Similarly, fig. 19-5 (B) shows the *acceptor level* at an energy slightly higher than the top of the valence band (with an energy difference again \sim several meV), which thereby accepts electrons from the valence band even at quite low temperatures. While the acceptor ions are all neutralized by the loosely bound holes at 0 K, almost all the holes become free at even low temperatures, which is expressed by saying that almost all the acceptor levels are occupied. The corresponding free holes in the valence band give a conductivity to the doped semiconductor.

The acceptor level thus corresponds to the energy of a negatively ionized acceptor atom residing at a lattice site, which is only slightly higher than the top of the valence band, which means that electrons from the valence band can be lifted to the acceptor level

with just a little supply of energy which can result from the thermal motion of the ions at even a very low non-zero temperature.

The Fermi level of the intrinsic semiconductor for each of the two types of semiconductors is also shown in fig. 19-5 (A), (B). In this figure, the temperature is assumed to lie in the range ~ 40 K to ~ 500 K - the range in which almost all the donor or acceptor atoms are ionized and, at the same time, appreciable thermal excitation from the valence band to the conduction band does not take place. For the n-type semiconductor, the Fermi energy lies just below the donor level while for the p-type semiconductor it lies just above the acceptor level.

In this context, recall our usage of the term Fermi energy (or level) whereby it may depend on the temperature, the dependence being a weak one for a semiconductor. Indeed, at 0 K, the Fermi energy of an n-type semiconductor lies *above* the donor level while that of a p-type semiconductor lies *below* the acceptor level. This is consistent with the fact that at 0 K, neither the donor atoms nor the acceptor atoms are ionized (recall that, in the case of the acceptor, an acceptor ion is neutralized by a bound hole at 0 K).

In the above introduction to doped semiconductors, I have emphasized on the role of the dopant atoms in generating the carriers - electrons in an n-type semiconductor and holes in a p-type semiconductor. However, a small number of carriers, of *both* signs are also generated independently of the dopant material by thermal excitation of electrons from the valence band to the conduction band - exactly as they are generated in an intrinsic semiconductor.

For an n-type semiconductor, the electrons generated in this manner are added to the pool of carriers arising from the dopant atoms. All these electrons taken together are known as the *majority carriers* in the n-type semiconductor, while the holes - few in number - generated by the intrinsic mechanism are referred to as the *minority carriers*. For a p-type semiconductor it is the other way round - holes are the majority carriers while electrons constitute the pool of minority carriers.

19.2.8 Intrinsic and doped semiconductors: summary

In summary, the electrical properties of intrinsic and doped semiconductors depend on the energy band structure in semiconductor crystals, of which germanium and silicon are common examples. An intrinsic semiconductor is a piece of pure crystal, for which the conduction band is empty at 0 K, and the energy gap between the valence and conduction bands is comparable to $k_B T$ (for $T \sim 300$ K, for instance). It has a low but non-zero conductivity, resulting from electrons and holes that act as negative and positive carriers respectively. These are generated by thermal excitation of electrons from the valence band to the conduction band. The conductivity of an intrinsic semiconductor increases with increase of temperature.

A doped semiconductor is obtained by adding a small quantity of some suitable impurity material, referred to as the dopant, to an intrinsic semiconductor. The dopant can be either a donor (group V element) or an acceptor (group III element), corresponding to an n-type or a p-type semiconductor respectively. The conductivity of a doped semiconductor is orders of magnitude higher than that of an intrinsic semiconductor, though the temperature dependence of conductivity is less pronounced.

The conductivity of a doped semiconductor is mainly due to the majority carriers resulting from the addition of the dopant atoms. In an n-type semiconductor the majority carriers are electrons, while a small number of holes also contribute to the conductivity as minority carriers. In a p-type semiconductor, on the other hand, the majority carriers are holes while a small number of electrons act as minority carriers.

19.3 The p-n junction diode

A piece of intrinsic or doped semiconductor crystal has only limited application in electronic circuits. A device in wide use, on the other hand, is a semiconductor crystal with *dissimilar doping* in two parts of it. It comes in several variant forms such as the ordinary diode (referred to as a p-n junction diode, a junction diode, or simply, a *diode* in brief), the Zener diode, and the light emitting diode, where these variants differ in their

structural and functional features as also in their circuit applications. In the present section I will outline the principles underlying the operation of the p-n junction diode and its use in *rectification*.

19.3.1 The junction diode: structural features

An overwhelming majority of semiconductor devices in use in electronic circuits employ semiconductor crystals with *inhomogeneous doping* in different parts or regions thereof. The junction diode, for instance, is made up of a semiconductor crystal (commonly, silicon) with one part doped with acceptor atoms and another with donor atoms. Fig. 19-6 shows this in a schematic way where the left half is an n-region and the right half a p-region.



Figure 19-6: A p-n junction diode (schematic); a rectangular geometry is shown for the sake of illustration; the interface between the p- and n-regions is shown with a line, being assumed to be infinitesimally thin; the two dots represent the two terminals of the diode, for connection to an external circuit.

While the figure depicts a rectangular geometry for the sake of illustration, in reality a cylindrical geometry is commonly used with one region enclosing the other.

This type of dissimilar doping in different regions of a semiconductor crystal can be achieved by one of several techniques in vogue. For instance, the acceptor atoms may be diffused into one region while donor atoms into another. The dopant may also be implanted into the semiconductor crystal in the form of ions that may be accelerated to a high velocity through an applied electric field and then made to hit the target region in the crystal.

The interface, i.e., the contact surface between the p- and the n-regions is of crucial importance in the functioning of the junction diode. In fig. 19-6 we have assumed that

the thickness of this interface is infinitesimal, though in reality a thickness spanning a number of inter-atomic bonds in the lattice is common.

In order to explain the functional characteristics of the junction diode, we first look at the situation prevailing inside the diode *at thermal equilibrium*.

19.3.2 The junction diode at thermal equilibrium

By thermal equilibrium I actually mean thermodynamic equilibrium where there is no macroscopic flow of any kind, i.e., flow of electrons or of holes in the present context. By contrast, things keep on happening all the while at a *microscopic* level, like electrons dancing around and getting scattered from vibrating atoms or ions. But none of this activity is discernible as a macroscopic flow because the macroscopic description takes a blurred view of things where the microscopic motions are averaged out. The imprint of the microscopic world remains only in the values of the averaged quantities of relevance characterizing a macroscopic system.

The basic fact to start from for a junction diode at equilibrium at a given temperature (say, T) is that the electrochemical potential (or the Fermi level, as we have loosely been referring to it) has to be *uniform throughout the volume of the diode*. For, a non-uniformity of the Fermi level would imply a flow of electrons (and holes) which is contrary to our assumption of equilibrium.

This gives us fig. 19-7 which is obtained by piecing together figures 19-5 (A) and (B) and placing their Fermi levels at the same energy in the vertical energy scale. The horizontal direction in the figure depicts change of position in the diode along an imaginary line passing through the contact surface between the p- and the n-regions. Looking away from the contact surface toward the left or the right gives us figures 19-5 (A) and (B) respectively, with the energy band structure of the n-type and the p-type semiconductor. It is *near* the contact surface that things look different because the bands have to *bend* in this region in order to join up. At the same time, the acceptor level in the p-region and the donor level in the n-region also get bent as shown.

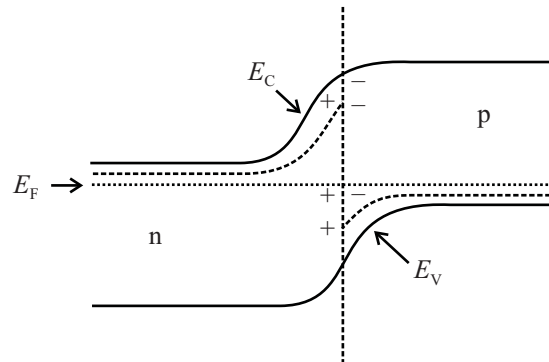


Figure 19-7: Energy levels in a p-n junction diode at thermal equilibrium (schematic); the Fermi energy (E_F) is the same in both the n- and p-regions; E_C and E_V denote the bottom of the conduction band and the top of the valence band respectively; the bands are bent in the region adjoining the interface, shown with a dashed vertical line; the donor and acceptor levels are also bent; the signs of charges appearing on the two sides of the interface are also shown.

One can imagine a p-n junction diode to be put together by joining up a p-type semiconductor with an n-type one, both at the same temperature, though this is not how the diode is actually fabricated. Immediately after the joining up, some transient activity will take place in the composite structure and it is only after the transient activity dies down that the situation depicted in fig. 19-7 comes to prevail.

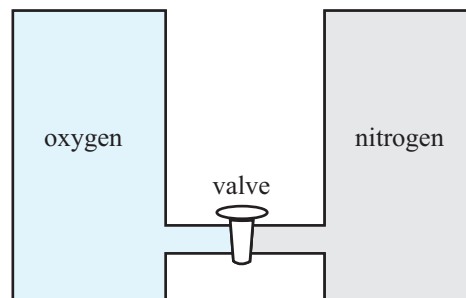


Figure 19-8: Analogy of two dissimilar gases diffusing across a valve; the valve is closed to start with; on opening the valve, there occurs a diffusion of the gases in opposite directions till equilibrium is achieved.

Here is an analogy. Figure 19-8 shows two cylinders containing two dissimilar gases (say, oxygen and nitrogen) at the same temperature and pressure, there being a valve in between which is initially closed. If now the valve is opened a transient flow will be generated through the valve in spite of the fact that the pressure and temperature are

the same on both sides of it. This is because the chemical potentials of oxygen and nitrogen are different on the two sides to start with, as a result of which *diffusion* will take place. Thermodynamic equilibrium prevails in the end as the process of diffusion comes to a stop.

Similarly, for the p-n junction diode, a diffusion of *electrons and holes* across the interface takes place before thermodynamic equilibrium is achieved. Recalling that there is a preponderance of electrons in the n-region and of holes in the p-region to start with, electrons will diffuse from the n- to the p-region while holes will diffuse in the opposite direction. However, unlike the case of the two dissimilar gases in the above analogy, the diffusion cannot continue till the concentrations of electrons and holes are equalized on the two sides, because these two diffusing components carry *charge*.

As a result of the diffusion of electrons and holes across the interface, negative and positive charges pile up in the p- and n-regions respectively, and consequently an *electric field* is set up in the region adjacent to the interface, *directed from the n-region to the p-region*. Another way of saying this is by referring to the variation of the electrical potential in the region adjacent to the interface, which looks as in fig. 19-9 (A) for a diode in thermal equilibrium. While the potential is uniform away from the interface, it is higher in the n-region and lower in the p-region. It varies significantly in the region adjacent to the interface, where positive and negative charge densities appear as shown in fig. 19-7.

This region close to the interface is termed the *depletion region*, or depletion layer - it extends on either side of the interface to a small distance, and its width depends on the degree of doping as also on the *bias* (see below) applied to the two terminals of the diode.

As the majority carriers (holes in the p-region and electrons in the n-region) diffuse in opposite directions across the interface, these recombine with oppositely charged carriers due to which a narrow region extending to both sides across the interface becomes *depleted* of mobile charges, and a space charge distribution is produced due to

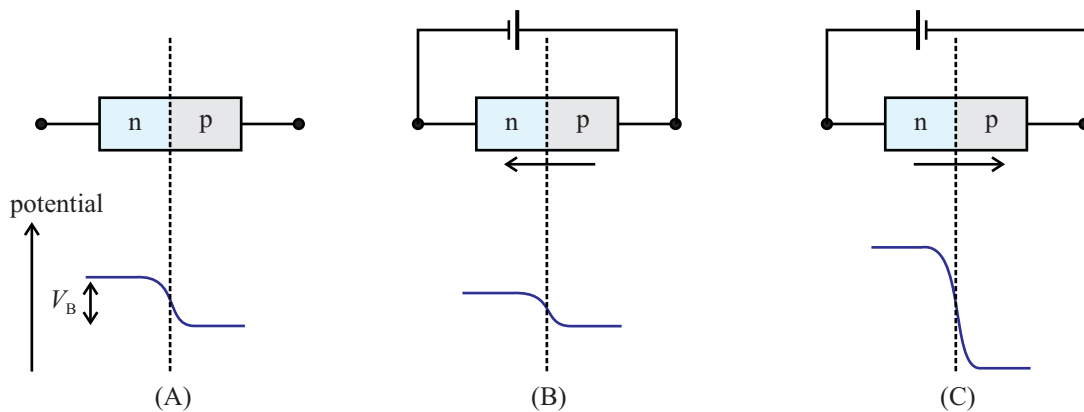


Figure 19-9: The junction diode in (A) thermal equilibrium, (B) forward bias and (C) reverse bias; in each case, the upper part of the figure represents the diode without or with bias, while the lower part shows the variation of electrical potential along a line directed from the n- to the p-region; in (A), V_B denotes the barrier potential, or barrier height for the unbiased diode; the barrier height is decreased for forward biasing while it increases in reverse bias.

the immobile ions - positively charged donor ions in the n-side and negatively charged acceptor ions in the p-side, added to which a small number of mobile carriers also remain. This depletion region acquires a high electrical resistance owing to the absence of the mobile carriers, and the space charge distribution results in an electric field in this region, directed from the n- to the p-side.

The potential difference between the interior of the n-region (i.e., away from the depletion layer) and the interior of the p-region is known as the *barrier potential* (also referred to as the barrier height) of the diode, and is denoted by V_B in fig. 19-9(A). An electron has to acquire an energy eV_B in order to cross over from the n- to the p-region, overcoming this barrier. A similar statement holds for holes crossing over from the p- to the n-region.

It is this potential difference between the n- and the p-regions of the diode that prevents an indefinite diffusive flow of electrons from the n- to the p-region and of holes from the p- to the n-region because it results in an *opposite* tendency of flow, known as *drift*, of electrons and holes. Equilibrium is achieved when the tendencies of diffusive flow and drift flow balance each other for each of the two types of carriers. This is precisely the condition that the Fermi energy be the same throughout the junction diode, including the depletion region as in fig. 19-7. The condition for zero net flow of holes and electrons

requires a particular space charge distribution and potential variation in the depletion region.

In summary, then, a junction diode in thermal equilibrium (at times referred to as an *unbiased diode*) develops a barrier potential difference (V_B) between the n- and the p-regions due to charges appearing on either side of the interface over a narrow region known as the depletion layer. No net flow of electrons or holes occurs in the diode in this equilibrium configuration - the current through the diode is zero.

It is important to distinguish between the electrical potential at various points within the diode and the electrochemical potential, the latter being relevant owing to the fact that there exists, in addition to the potential gradient, a concentration gradient of electrons and holes across the diode. It is the electrochemical potential that determines the net motive force driving the flow of the electrons and holes. At equilibrium, the electrochemical potential has to be uniform throughout the diode. This explains why it is not possible to measure the barrier potential with a voltmeter since no current would flow through the voltmeter if it were joined across the diode.

19.3.3 The junction diode in forward and reverse bias

The term *bias* with reference to an electronic device usually means appropriate DC voltages or currents applied to the terminals of the device so that it can perform some desired function in a circuit. For a junction diode, it means a DC voltage applied between the p- and n-terminals of the diode. This is shown in fig. 19-9 (B) and (C), where a DC voltage source is connected to the terminals of the diode. While an actual circuit usually includes other circuit elements as well, the figure gives us the essential point in diode biasing.

Thus, in fig. 19-9 (B), the negative terminal of a DC voltage source (represented by the short vertical line) is connected to the n-terminal of the diode while the positive terminal of the source (longer vertical line) is connected to the p-terminal of the diode. This is known as *forward biasing* while, in (C), the connection is made with opposite polarity,

corresponding to *reverse biasing* of the diode.

As shown in fig. 19-9 (B), forward biasing leads to a decrease in barrier height because it is applied in opposition to the potential difference existing between the n- and the p-regions of the unbiased diode. The diode is no longer in an equilibrium state because, as a result of the lowered potential barrier and hence a weaker electric field in the depletion region, the tendency of drift motion of holes and electrons gets diminished while, on the other hand, the tendency of diffusion increases, the latter because of an increased concentration gradient of the carriers resulting from a narrowing of the depletion region caused by the lowered potential barrier. A net current is thereby set up through the diode, directed from the p- to the n-region (direction of arrow). This is a substantial current since it is mostly caused by *majority carriers* in the n- and p-regions on the two sides of the barrier, and it increases rapidly with the bias voltage between the two terminals of the diode.

In fig. 19-9 (C), on the other hand, the reverse polarity of the bias voltage increases the height of the potential barrier, thereby making stronger the tendency of drift flow of the majority carriers (electrons from the n- to the p-region and holes in the reverse direction) while, at the same time, the width of the depletion region increases, thereby reducing the concentration gradient across the barrier and weakening the tendency of diffusive flow. Thus, once again, the balance between the two tendencies is upset in favour of the drift flow, and a current is set up, now directed *from* the n- *to* the p-region. This is referred to as the *reverse current*. It does not change much with the applied reverse bias voltage between the terminals of the diode, and has a rather strong temperature dependence. It is mostly caused by the minority carriers in the two regions, where we have already seen that their production mechanism depends on thermal excitations of the electrons from the valence to the conduction band, which explains the temperature dependence of the reverse current.

The correct description of the flow of holes and electrons through a diode involves quite subtle considerations, most of which I have had to gloss over in this brief and introductory presentation. For instance, the flow characteristics of majority and minority carriers in the n- and p-sides of the depletion region differ markedly

from the corresponding features in the n- and p-type bulk regions. What I have attempted here is nothing more than a plausibility argument that may help you find an explanation, in a manner of speaking, of the I - V characteristic described in sec. 19.3.4.

19.3.4 Junction diode: current-voltage graph

All the operational features of the junction diode at any given temperature can be summarized in a single graph referred to as the current-voltage characteristic of the diode. It relates the current I through the diode to the voltage V between the diode terminals. I is defined as the current flowing from the p- to the n-terminal. If, in an observation, the current is found to flow from the n- to the p-side then that would mean that I is negative for that observation. Similarly, V is defined as the potential difference between the p- and the n-terminal. It can be negative if the n-terminal happens to be at a higher potential compared to the p-terminal, as in fig. 19-9 (C).

For a typical junction diode the graph (referred to as its I - V characteristic) looks like fig. 19-10. It is a sharply bent graph, quite distinct from a straight line graph that one would obtain for a simple resistor. The latter is an ohmic element while the bent look of the I - V graph of the junction diode tells us that it is a non-ohmic element.

The portion of the graph with $V > 0$, $I > 0$ corresponds to forward bias while that with $V < 0$, $I < 0$ corresponds to reverse bias. The former includes a steeply rising portion that starts from $V \sim 0.6$ V for a silicon diode. This is known as the *cut-in voltage* and corresponds to the *knee* of the bent graph. For reverse bias, the current is negative and of small magnitude. When the reverse bias voltage is relatively high, the magnitude of the reverse current becomes constant (I_0 in the figure), and is referred to as the *reverse saturation current*.

The figure shows a point P on the graph, for which the voltage and current are given by the abscissa and ordinate of P. The ratio of the two gives what is known as the *DC resistance* of the diode for the point P. At the same time, the narrow triangle constructed around P gives the *incremental voltage* δV and *incremental current* δI , which are small changes in voltage and current around P. The ratio of these two gives the *AC resistance*

of the diode for the point P.

As the point P moves up the steeply rising part of the graph, the DC resistance decreases and, at the same time the AC resistance also decreases - more rapidly than the DC resistance. For a forward voltage even slightly higher than the cut-in voltage the DC and AC resistances become very small, the former $\sim 10\Omega$ and the latter as low as $\sim 1\Omega$ for a typical diode. The two resistances are important operational characteristics of the diode depending on whether one is looking at DC currents and voltages or AC ones in a circuit containing the diode.

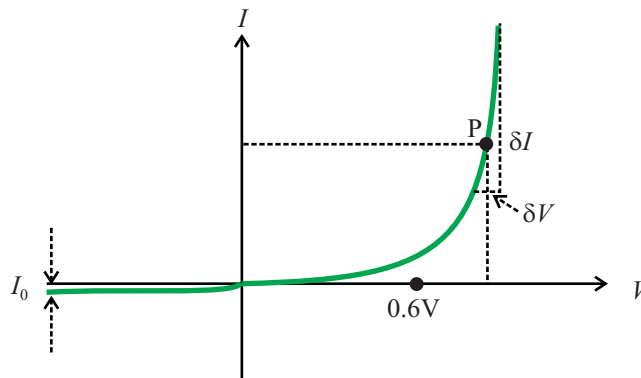


Figure 19-10: Current-voltage graph of a junction diode; for a point P on the graph, the voltage and current are given by the abscissa and ordinate of P; their ratio gives the DC resistance of the diode for the point P; at the same time, the incremental voltage δV and incremental current δI are shown; their ratio gives the AC resistance; in the graph, the part lying to the right of the current axis corresponds to forward bias while the one lying to the left corresponds to reverse bias; I_0 is known as the reverse saturation current; for a silicon diode, the steeply rising portion starts from $V \sim 0.6V$.

For reverse bias, on the other hand, the DC and AC resistances, defined similarly, turn out to be *very large*, even as large as $\sim 10^6 \Omega$ or still larger.

If the reverse bias voltage of a diode is made large in magnitude (commonly, of the order of 50 V or more), the diode *breaks down* because of a large number of electrons being pulled out of the valence band by the reverse electric field in the depletion layer and raised to the conduction band, when the carrier concentration (electrons in the conduction band and holes left in the valence band) increases greatly. As a result the

reverse current through the diode increases dramatically (not shown in fig. 19-10). In regular diode operation, however, this break-down condition is avoided by limiting the magnitude of the reverse voltage to relatively low values.

19.3.5 Junction diode: summary

In summary, a junction diode is made up of a p-region and an n-region, with a thin interface between the two. With no external voltage applied to the two terminals of the diode, a state of equilibrium prevails for which the current through the diode is zero. However, a potential difference V_B exists between the p- and n-regions, the variation of the electrical potential occurring mainly over a thin layer around the interface, referred to as the depletion layer. Incidentally, this potential difference cannot be detected or measured by connecting a voltmeter between the two terminals of the diode because, with no external voltage applied to the terminals of the diode, no current would flow through the circuit (because the electrochemical potential is uniform throughout the diode; refer to the I - V graph of fig. 19-10) and the voltmeter would show no deflection. A uniform potential at all points in the external circuit, which has no source of EMF in it, is achieved by means of contact potentials being developed at the two points of contact with the diode.

The diode can be operated with either a forward or a reverse bias voltage applied to it as in fig. 19-9 (B), (C). The operational characteristics of the diode are expressed by the I - V graph of fig. 19-10. In forward bias, especially for the voltage above the cut-in value, the AC and DC resistances of the diode are very small, while these are *large* in reverse bias.

In other words, the diode is a device that has a low resistance *or* a high resistance depending on the voltage applied to it. It can be made to *switch* from the low to the high resistance mode by changing the applied bias voltage. Another way of expressing the same thing is to say that the diode acts as a voltage-controlled *switch*.

An electrical switch is a device that can be operated in two modes - *on* or closed when it allows current to flow in a circuit, and *off* or open when the current is stopped. While household electrical switches are operated manually, a diode is made on or off with the

voltage applied between its terminals. With a forward bias voltage, the diode offers a low resistance (almost zero), and allows current to flow - the switch is *on*. With a reverse bias voltage on the other hand, the diode resistance is very high, and the switch is *off*.

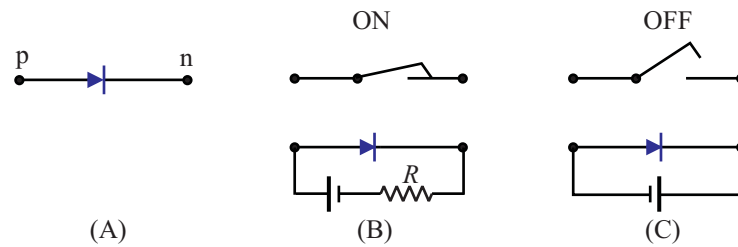


Figure 19-11: The junction diode as a voltage-controlled switch: (A) circuit symbol of a diode - the dot on the side of the arrowhead denotes the p-terminal and the one on the side of the short vertical line the n-terminal; (B) diode is forward biased (compare with fig. 19-9 (B)), when it resembles a switch in the *on* state, the latter being shown symbolically on top of the diode; (C) diode is reverse biased (see fig. 19-9 (C)), resembling a switch in the *off* state: in (B), R denotes a resistance inserted in the circuit for the purpose of controlling the current through the diode.

Figure 19-11 (A) shows the *circuit symbol* of a p-n junction diode where the p- and n-terminals are indicated. In the same figure, (B) and (C) illustrate the operation of the diode as a *voltage controlled electrical switch*, which is *on* in forward bias and *off* in reverse bias. Another way of expressing the operational characteristics of the diode is to say that it is a voltage controlled *resistance* - low in forward bias and high in reverse bias.

19.3.6 The diode as rectifier

The fact that the diode is capable of operating in one of two modes - the forward biased *on* mode (low resistance), and the reverse biased *off* mode (high resistance) - makes it suitable for a good number of applications. One of these involves *rectification* of an AC voltage into a unidirectional one.

Fig. 19-12 shows the basic rectifier circuit. In this figure, as also in figures depicting other electronic circuits, the hatching with horizontal lines depicts the 'ground', a common junction in the circuit to which several terminals are connected (this results in economy in connection) and which is commonly *earthed* so as to minimize the risk of

electrical shock to one working with the circuit (most circuits need *power supplies* - DC or AC voltage sources that require connection to the supply mains). The circle with the wavy line drawn in indicates an AC voltage source, in the present instance the source whose output is to be rectified. The figure makes use of the circuit symbol of a p-n junction diode (fig. 19-11(A)).

In the circuit diagram, R denotes a resistor (more precisely, the *resistance* of the resistor), usually of the order of a $k\Omega$, used to limit the current through the diode when it is in the *on* mode (an excessive current through the diode may heat it up and damage it). The output voltage of the set-up appears across this resistance, commonly referred to as a *load* resistance.

In the figure, (A) and (B) denote the input and output voltage waveforms (a pictorial depiction of the variation of the voltage with time). The input waveform is made up of successive cycles, where each cycle in turn is split into two half-cycles with, respectively, *positive* and *negative* values of the voltage - the characteristic feature of an AC voltage.

During the positive half-cycle, the p-terminal of the diode is at a higher potential compared to the ground, resulting in the diode being *on*. A current flows through the circuit and the input voltage appears across R (with only a small part of the voltage appearing across the diode itself). This accounts for the output waveform during the positive half-cycle (two such half-cycles are shown). Incidentally, the '+' and '-' symbols in the figure are used only to *define* the voltages V_{in} and V_{out} - for instance V_{in} is defined as the instantaneous voltage at the p-terminal of the diode relative to the voltage at the ground terminal. These symbols have nothing to do with the signs of the actual values assumed by V_{in} and V_{out} at any given instant.

During a negative half-cycle, on the other hand, the p-terminal of the diode is at a lower voltage with respect to the n-terminal, and the diode is *off*. No current flows through the circuit, and the output voltage across R is almost *zero*. This accounts for the horizontal parts in the output waveform during the negative half-cycles of the input.

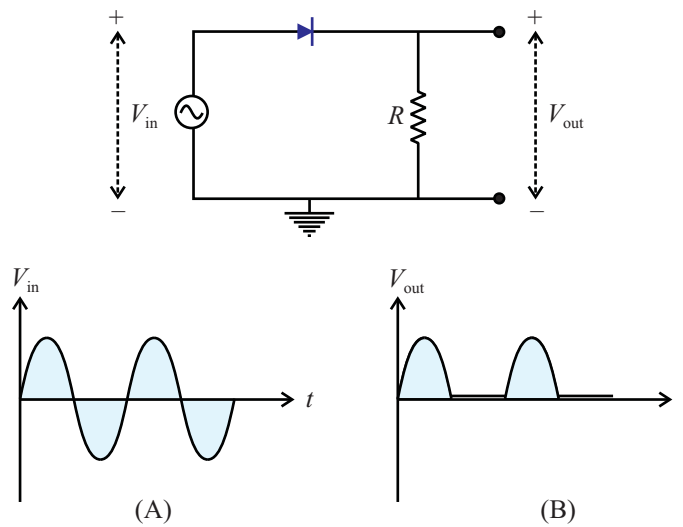


Figure 19-12: The basic circuit of a rectifier, together with its input and output waveforms; (top) the rectifier circuit, where a p-n junction diode is represented by its circuit symbol; an AC voltage source is connected between the p-terminal of the diode and the *ground* terminal; R denotes a resistance across which the output voltage appears; V_{in} and V_{out} denote the input and output voltages at any given instant of time; (below) (A) the input, and (B) the output waveforms; two cycles are shown; each cycle of the input waveform consists of a positive and a negative half-cycle; the negative half-cycles are cut-off by the set-up so that the output voltage is zero during these half-cycles.

The set-up has thereby achieved rectification - cutting off of the negative parts of an input waveform.

A diode in a rectifier circuit thus acts effectively as a switch, shutting off the negative half of a full cycle of an alternating voltage presented to it. More specialized high speed switching action of diodes is made use of in *power control* circuits.

Problem 19-1

The current through a p-n junction diode is $I = 30$ mA when the voltage across it is $V = 0.55$ V, while for voltages 0.6 V and 0.65 V, the current is seen to be 40 mA and 55 mA respectively. Obtain the DC and AC resistance of the diode at 0.6V.

Answer to Problem 19-1

HINT: The DC resistance at $V = 0.6$ V is $r_{DC} = \frac{V}{I} = \frac{0.6}{40 \times 10^{-3}}$ Ω , i.e., 15 Ω . As for the AC resistance,

the variations in current and voltage at a mean voltage of 0.6V are $\delta I = 25 \text{ mA}$ and $\delta V = 0.1 \text{ V}$ respectively. Hence $r_{AC} = \frac{\delta V}{\delta I} = \frac{0.1}{25 \times 10^{-3}} \Omega$, i.e., 4Ω .

Problem 19-2

In fig. 19-11(B), a 1.5 V DC source is used and the resistance R is 18Ω , for which the current through the circuit is 50 mA. Calculate the DC resistance offered by the diode and the power dissipated in it.

Answer to Problem 19-2

HINT: The voltage drop across R being $50 \times 10^{-3} \times 18 \text{ V}$, i.e., 0.9 V, that across the diode is $(1.5 - 0.9) \text{ V}$, i.e., 0.6 V. Thus, the DC resistance of the diode under the given operating conditions is $\frac{0.6}{50 \times 10^{-3}} \Omega$, i.e., 12Ω . The power dissipated in the diode is the product of the voltage drop across it and the current, i.e., $0.6 \times 50 \times 10^{-3} \text{ W}$, i.e., 30 mW (the power supplied by the DC source is 75 mW, of which 45 mW is dissipated in the resistance R).

Problem 19-3

In the rectifier circuit of fig. 19-12, the supply voltage from the AC source is $V(t) = V_0 \sin \omega t$, with $V_0 = 20 \text{ V}$ and $\omega = 1 \text{ kHz}$. If the load resistance is $R = 10 \text{ k}\Omega$, find the power dissipated in the load, assuming ideal operating conditions. What is the peak forward current through the diode?

Answer to Problem 19-3

HINT: The term 'ideal operating conditions' is to be interpreted here as implying zero output resistance of the AC source, zero resistance of the diode under forward bias, and infinite resistance under reverse bias. Hence, considering one complete cycle of the input voltage from $t = 0$ to $T = \frac{2\pi}{\omega}$, the output voltage across the load will be $V_{\text{out}(t)} = V_0 \sin \omega t$ during the positive half cycle ($0 \leq t \leq \frac{\pi}{\omega}$), and $V_{\text{out}(t)} = 0$ during the negative half cycle ($\frac{\pi}{\omega} \leq t \leq \frac{2\pi}{\omega}$), the corresponding currents being $I_{\text{out}}(t) = \frac{V_0}{R} \sin \omega t$ during the positive half cycle, and 0 during the negative half cycle. The power dissipated in the load is (refer to formula (13-73), in which the contribution to the integral for the negative half cycle will be zero in the present context)

is $\mathcal{P} = \frac{1}{T} \int_0^T V_{\text{out}}(t) I_{\text{out}}(t) dt = \frac{V_0^2}{RT} \int_0^T \sin^2 \omega t dt = \frac{V_0^2}{4R} = 10 \text{ mW}$. The peak forward current is $I_0 = \frac{V_0}{R} = 2\text{mA}$.

19.3.7 Special-purpose diodes

The p-n junction diode described above is the commonly used form of the semiconductor diode, while a number of special purpose diodes are also used in a large number of applications. All of these are based on the same fundamental principles as the regular p-n junction diode but have some additional constructional features or modifications built into them so as to make them function in novel ways. I will introduce below three of these special purpose diodes currently in wide use.

19.3.7.1 The Zener diode

At the end of sec. 19.3.4, I referred to the break-down condition for the p-n junction diode which occurs when the reverse voltage applied to the diode is made to increase to a high value, commonly not realized in regular diode operation.

The *Zener diode*, however, is a special purpose diode where the diode is *made to break down* by applying an appropriate reverse voltage since this feature of break-down is made use of in a number of circuit applications. For this purpose, the *doping level* (i.e., the donor and acceptor concentrations in the n- and the p-regions respectively) of the diode is made high, which results in a reduction of the breakdown voltage by facilitating the process of tearing away of electrons from the valence band to the conduction band.

As the doping level is made high, the depletion layer at the p-n junction becomes thinner and the electric field intensity in the depletion layer for any given reverse voltage increases correspondingly. It is this increased electric field intensity in the depletion region that is responsible for the facilitated break-down of the diode.

Fig. 19-13 shows the I - V characteristic of the Zener diode including the break-down region (marked AB in the figure) where it is seen that the reverse current increases

steeply with almost no increase in the reverse voltage. In other words, when the diode breaks down, the reverse current through it can be made to increase while the reverse voltage across the diode remains at the constant magnitude V_B . In the figure, I_Z denotes the Zener current and V_Z the Zener voltage so that, in the break-down region, $V_Z = -V_B$.

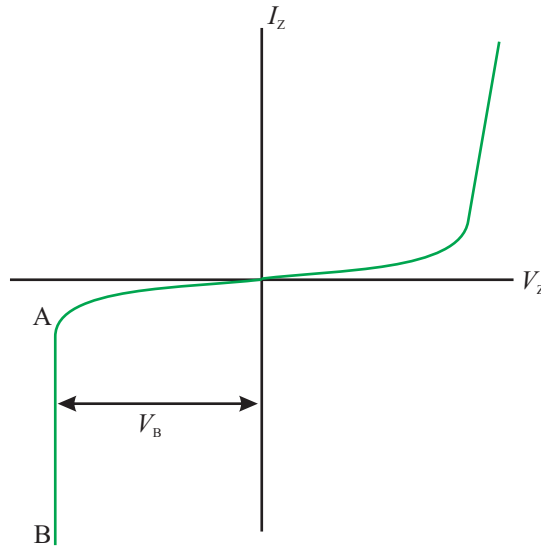


Figure 19-13: I - V characteristics of a Zener diode; the break-down region AB is indicated in which the Zener current changes as the Zener voltage remains constant at the value $-V_B$.

Fig. 19-14(A) shows the circuit symbol of the Zener diode which differs from that of a regular p-n junction diode (see fig. 19-11(A)) in that the n-terminal of the diode is represented by a wavy line instead of a straight segment. Fig. 19-14(B) on the other hand, shows a simple application of the Zener diode as a *voltage regulator*. In order to understand what is meant by voltage regulation, look at the simple circuit of fig. 19-14(C), in which a resistance R is connected across a DC voltage source of EMF E and internal resistance r (shown explicitly in the circuit). The voltage across the resistance R in this circuit is seen to be

$$V = E - Ir = \frac{ER}{r + R}, \quad (19-1)$$

where I is the current drawn from the voltage source. One observes that for a given source, the voltage across the resistance R (commonly referred to as the *load* resistance)

decreases as R is made to decrease or, equivalently as the current I (the *load* current or, in brief, the *load*) is made to increase. This dependence of V on R or on I is enhanced for a source of higher internal resistance.

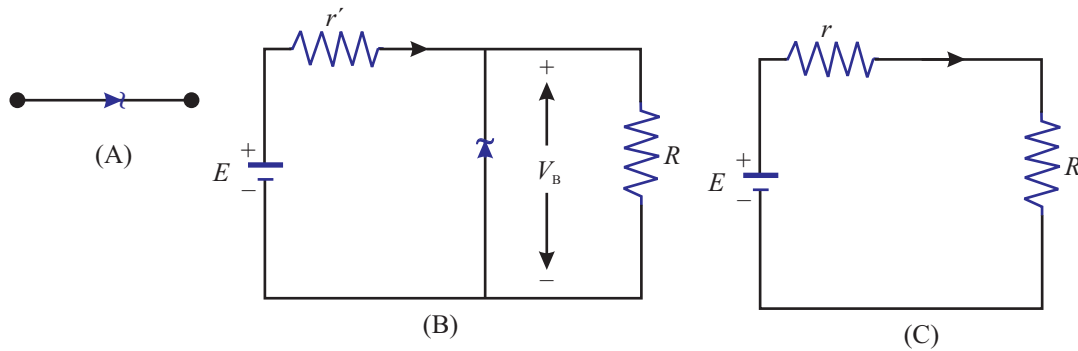


Figure 19-14: (A) the circuit symbol of Zener diode; (B) a simple voltage regulator circuit using a Zener diode; the functioning of this circuit is to be understood by comparing it with the circuit in (C); (C) a load resistance R connected directly to a source of DC voltage, of internal resistance r and EMF E .

For many applications, however, one needs a *constant voltage* to appear across the load resistance independent of the load current. This is achieved in the circuit of fig. 19-14(B) where the Zener diode is connected in parallel across the load resistance, and where the resistance r' may include, in addition to the internal resistance of the voltage source, an appropriate extra resistance put in for the proper operation of the circuit. When appropriately designed, the circuit operates with the Zener diode in the breakdown condition, as a result of which the voltage across R is V_B , i.e., a constant voltage regardless of the value of R or of the current drawn from the source. This is, in basic terms, what is meant by the term 'voltage regulation'.

While the circuit of fig. 19-14(B) illustrates the basic principle of voltage regulation using a Zener diode, an actual voltage regulator circuit often requires a number of other components and involves more elaborate considerations for its design.

Zener diodes are used in a great many circuit designs, where it diode serves the basic purpose of maintaining a constant voltage between any two chosen points in the circuit.

Problem 19-4

In the circuit of fig. 19-14(B), a 18 V DC source is used and the break-down voltage of the Zener diode is 10 V. If the current limiting resistance r' is $100\ \Omega$ and the load resistance is $R = 200\ \Omega$, find the current through the Zener diode, and the power dissipated in it.

Answer to Problem 19-4

The voltage drop across the resistance r' is $(18 - 10)$ V, i.e., 8 V, and hence the current supplied by the source is $\frac{8}{100}$ A, i.e., 80 mA. The load resistance R being connected in parallel to the Zener diode, the voltage across it is 10 V (the supply voltage being larger than the break-down voltage of the Zener diode, the latter operates in the break-down mode and the load voltage equals the break-down voltage; the fact that the load voltage is independent of the value of R (within limits) is responsible for the voltage regulating action of the Zener diode) and hence the current through it is $\frac{10}{200}$ A, i.e., 50 mA. Hence the current through the Zener diode is $(80 - 50)$ mA, i.e., 30 mA (the current supplied by the source is divided into two parts; the reverse current through the Zener diode, of magnitude 30 mA in the present instance, is responsible for the diode operating in the break-down mode). Hence the power dissipated in the Zener diode is $P = 10 \times 30$ mW, i.e., 300 mW.

19.3.7.2 The light emitting diode

The *light emitting diode* (LED in brief) is another special purpose diode that emits light when carrying a forward current. The p- and n-regions of this diode are made of alloys of group III and group V elements, like Ga-As and Ga-P, where the two elements are mixed in appropriate proportions. As a forward current flows through the diode, the electrons injected from the n- to the p-region of the diode combine with holes in the p-region, thereby making a transition from the conduction band in the p-region to the valence band. In the process, a certain amount of energy is released by each electron (a corresponding process takes place with holes injected from the p- to the n-region).

The materials of which the LED's are made of (such as Ga-As and Ga-P) belong to the class of *direct bandgap semiconductors* and have the special property that this energy

comes out in the form of visible light rather than in the form of thermal vibrations of the atoms making up the crystalline structure (as in the case of the *indirect bandgap* semiconductors like germanium and silicon). Typically, an LED is enclosed in a glass casing that acts as a lens concentrating the light rays coming out of the LED.

The distinction between direct and indirect bandgap semiconductors relates to details of their band structures, especially to the way that the energies of the electrons and holes in the conduction and valence bands respectively vary with their momentum, the latter being, strictly speaking, a quantum mechanical concept, analogous to that of the quantum mechanical momentum of a free particle. The requirements of energy and momentum conservation in the process of combination of an electron and a hole (this is referred to as one of *recombination* in the context of a semiconductor device) do not allow the emission of an optical photon in the case of an indirect bandgap semiconductor while, for a direct bandgap material, the emission of an optical photon is compatible with the conservation principles.

There exists a range of materials that can be used for the fabrication of LED's, corresponding to a range of colors of the light emitted by these diodes.



Figure 19-15: The circuit symbol of (A) a light emitting diode and (B) a laser diode.

LED's are rugged and useful devices with low power consumption and high light emitting efficiency, and are destined to replace incandescent lamps and gas discharge tubes in home applications. LED's are currently in wide use for the purpose of indicating voltages at various points in electronic and electrical circuits, as also for display purposes.

Fig. 19-15(A) depicts the circuit symbol of a LED.

19.3.7.3 The laser diode

A laser diode is, in principle, a *heavily doped (degenerate doping)* light emitting diode in which conditions for *stimulated emission* are realized, resulting in the production of an intense and coherent beam of light known as a *laser beam*.

As in the light emitting diode, photons (energy quanta of electromagnetic radiation, see chapter 16 for an explanation) are emitted when the diode carries a forward current, but now the process of recombination of electrons and holes (electrons dropping down from the conduction band to the valence band so as to combine with holes with a release of energy, and a corresponding process for holes) is made to occur in an extremely small region specially designed for the purpose so that the photons remain in close proximity to the electrons and holes participating in the process and these photons, in turn, influence the process of further photon emission by recombination. This is referred to as the process of *stimulated emission* which differs fundamentally from *spontaneous emission*, i.e., in the context of the diode, emission by recombination unaided by photons in the neighborhood of the region where the recombination takes place, the latter being the dominant process in an ordinary LED.

The process of stimulated emission, explained in greater details in sec. 15.6.4, makes possible a *cascading* of emission events since the photons emitted in a stimulated emission may produce further events of stimulated emission, and the process snowballs. This leads to what is referred to as *light amplification*, which is made possible by a *resonating optical cavity* built into the diode. The cavity is essentially made of a pair of highly reflecting parallel mirrors which cause the photons to be reflected back and forth between them, producing a *standing wave* of the light. The standing wave corresponds to a sharply defined frequency since only such a frequency resonates with the cavity, being selected out by virtue of the cavity dimensions. Further, the photons in the standing wave all have a common direction of propagation since photons propagating in other directions get lost from the cavity.

1. As explained in section 15.6 (see, in particular, sec. 15.6.6), a *population inversion* is necessary for lasing action to become possible. In the case of the laser diode,

the condition of population inversion is achieved by manipulating the 'quasi Fermi levels' in the n- and p-regions of the diode so that these are, respectively, above the bottom of the conduction band and below the top of the valence band in the forward biased condition, when light emission takes place. In this context, refer to fig. 19-5 where the Fermi levels are seen to be located below the conduction band and above the valence band respectively, this being the configuration in an unbiased diode. By appropriately manipulating the doping (degenerate doping) and the forward biasing, the required configuration of the Fermi levels (these are referred as the *quasi* Fermi levels when the equilibrium configurations of the energy bands in the n- and p-regions are disturbed) is achieved, and lasing action occurs inside the depletion region.

2. A laser diode is made to operate with a large forward current so as to make possible the realization of the conditions necessary for light amplification. This requires a careful control of the forward bias voltage, which is achieved with special arrangements.

When the standing wave attains a certain amplitude by the cumulative addition of photons into it, an intense, highly collimated, and monochromatic beam of photons comes out of the cavity in virtue of the small transmissivity (made possible by appropriate design) of one of the two reflecting mirrors.

While this constitutes the basic principle underlying the laser diode in its barest outline, the actual details of construction and operation constitute a subject in itself. Laser diodes of numerous different constructional and operational types are now available for a wide variety of applications.

Some of the more common applications of laser diodes include those in bar-code readers, image scanners, CD players, CD-ROMs, in DVD technology, in fiber-optic communications, and in laser pointers. The circuit symbol of a laser diode is shown in fig. 19-15(B).

19.4 The bipolar junction transistor

The p-n junction diode is useful because of its sharply bent I - V characteristics, due to which it acts like a voltage-controlled resistance. However, apart from the externally applied voltage to the diode, there is no other controlling parameter affecting the operation of the diode. The *bipolar junction transistor* (BJT in short; also commonly referred to as, simply, the *transistor*) is an electronic device whose operation involves more than one controlling parameters, as a result of which it has a much wider range of applicability.

The BJT is a three-terminal semiconductor device made up of alternating layers of n-type and p-type semiconductors. However, it is not the only type of transistor in common use, other transistor types of equally varied applicability being the JFET (*junction field effect transistor*) and the MOSFET (*metal oxide semiconductor field effect transistor*). In this introductory presentation, however, I will confine myself to an elementary survey of the BJT.

19.4.1 The emitter, the base, and the collector

The BJT comes in two varieties, the n - p - n and the p - n - p transistors. If you look inside an n-p-n transistor, for instance, you will find three doped regions, namely, two n-type regions on either side of a p-type region. A p-n-p transistor similarly involves two p-type regions separated by an n-type region. Fig. 19-16(A) is a schematic representation of an n-p-n transistor, where an external terminal connects to each of the regions. These regions along with the external terminals of the device, are referred to as the *emitter* (E), *base* (B), and the *collector* (C) respectively. Though the three regions are shown in the figure to be of rectangular shape, it is common for these to have a cylindrical geometry, the innermost cylindrical region being the n-type emitter, surrounded by two other cylindrical regions, the p-type base and the n-type collector, the collector being the outermost region. The interface between the emitter and the base has therefore a smaller area compared to that between the base and the collector.

This is important for the reason that the currents crossing these interfaces are, in general, almost equal while the *voltages* across these are markedly different, the emitter-

base voltage being usually much smaller compared to the base-collector voltage. As a result, the rate of heating (current times voltage) in the former is much smaller than the latter, requiring a larger area for the latter in order that the heat may be dissipated and the device does not get unduly heated up.

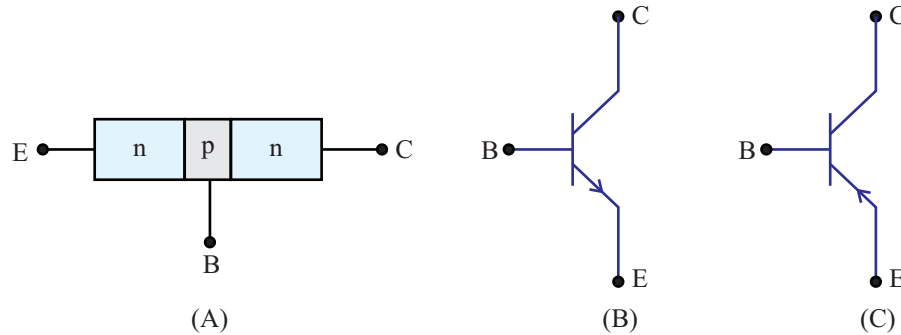


Figure 19-16: (A) Schematic representation of the three regions in an n-p-n transistor - each of the three regions is connected to an external terminal; E, B, and C are the terminals respectively to the emitter, the base, and the collector; the base is a thin region separating the emitter and the collector, the doping level in it being low; the doping level is the highest in the emitter, followed by that in the collector; for a p-n-p transistor the three regions (E, B, and C) correspond to p-type, n-type, and p-type doping respectively; (B) and (C), circuit symbols for an n-p-n type and a p-n-p type transistor respectively; the direction of the arrow on the oblique arm connecting with E differs in the two cases.

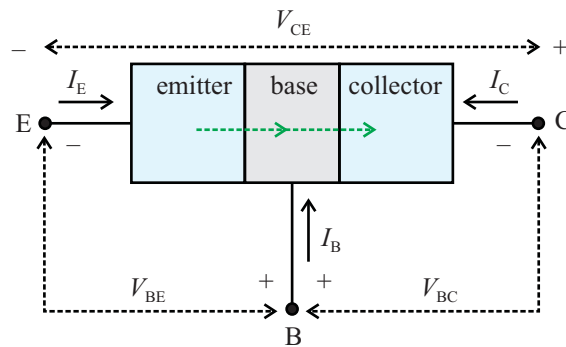


Figure 19-17: Defining the currents at the terminals and the voltages between pairs of terminals of a transistor; the currents are all defined to be ones flowing into the respective terminals as indicated by the arrowheads, which have nothing to do with the actual directions of these currents; the potential differences are defined with reference to the '+' and '-' symbols, which likewise have nothing to do with the signs of the actual values of these voltages; the flow of electrons from the emitter to the base and the one from the base to the collector (an n-p-n transistor is shown for the sake of concreteness), with the transistor operating in the *active* mode (see sec. 19.4.4), are depicted with dotted arrows.

One other important constructional feature of the BJT is that the *doping level* in the emitter region is usually appreciably higher than that in the collector region, the doping level in the base region being the lowest of the three. Finally, one other feature of crucial significance is that the width of the base region is very small compared to the emitter and collector regions, i.e., in other words, the base is a thin layer separating the emitter and the collector regions.

The circuit symbol of an n-p-n transistor is given in fig. 19-16(B), while the symbol for a p-n-p transistor is given in fig. 19-16(C). The three terminals (E, B, and C) are shown along with each symbol.

19.4.2 The two-diode model of the transistor

The operation of the transistor can be conveniently described and understood by looking at it as a combination of two junction diodes - the base-emitter (B-E) diode and the base-collector (B-C) diode, each of these two diodes acting like a high resistance when reverse biased and a low resistance when forward biased, regardless of the other diode. Depending on how the two diodes are biased, the transistor can be made to operate in one of three *modes* - the *active* mode, the *saturation* mode, and the *cut-off* mode. These three modes correspond to distinct graphical relations between the currents at the three terminals of the transistor and the voltages across the three regions.

19.4.3 Transistor currents and voltages

Fig. 19-17 shows, in a schematic representation of an n-p-n transistor, the various currents and voltages at and between the terminals of the transistor. I_E , I_B and I_C are the currents at the emitter, base, and the collector respectively, all defined as currents flowing *into* the respective terminals, as illustrated with the arrows. If, for instance, a current measuring 2 mA happens to flow *out* of the emitter terminal, then that would mean $I_E = -2$ mA.

Similarly V_{BE} , V_{BC} , and V_{CE} denote the potential difference between the base and emit-

ter, that between the base and collector, and that between the collector and emitter respectively. The '+' and '-' symbols are used to illustrate these definitions, and do not represent the signs of the actual values of the respective voltages. Thus, for instance, a value $V_{BC} = -11.2$ V means that the base is at a *negative* voltage with respect to the collector.

19.4.4 The transistor in the active mode

Consider now an n-p-n transistor (I refer to the n-p-n transistor for the sake of concreteness; corresponding statements hold for a p-n-p transistor with all currents and voltages imagined to be reversed), with its B-E diode *on* (forward biased) and its B-C diode *off* (reverse biased). Accordingly, V_{BE} is positive and V_{BC} is negative, while $V_{CE}(= V_{BE} - V_{BC})$ is also positive.

Looking back to fig. 19-10, one observes that the value of V_{BE} has to be close to 0.6-0.7 V (for a silicon transistor) for the B-E diode to be on - a lower voltage would effectively turn the diode off, and even a slightly higher voltage would cause a steep rise in the diode current causing the device to get damaged. We will assume that, with the B-E diode on, $V_{BE} \approx 0.7$ V for a silicon transistor, with only small variations during the actual operation of the device.

The forward biasing of the B-E diode causes a stream of electrons to be injected from the emitter into the base (dotted arrow in fig. 19-17) where they move on to the B-C interface. While moving through the base region a few of these electrons *recombine* with the holes in the base. Since, however, the width of the base is small and the doping level in it is low, these recombinations are rare, and most of the electrons injected from the emitter into the base reach the B-C junction. On reaching the B-C junction, these electrons are drawn into the collector region (the second dotted arrow in the figure) since, the B-C diode being reverse biased, the collector is at a higher voltage with reference to the base.

In order to maintain electrical neutrality in the emitter region in the face of a stream of electrons being injected from the emitter into the base, electrons flow from the external circuit into the emitter, causing an emitter current I_E to be set up, which is *negative*

for an n-p-n transistor in the active mode. Similarly, the electrons moving into the collector find their way into the external circuit through the collector terminal, causing a collector current I_C to be set up, which is *positive* for an n-p-n transistor in the active mode where, moreover,

$$I_C \approx -I_E. \quad (19-2)$$

The small number of holes recombining with the migrant electrons in the base region is made up for by electrons leaving the base region into the external circuit, resulting in a base current I_B , which is positive for an n-p-n transistor in the active mode. The charge neutrality for the entire device implies that the three currents I_E , I_B , and I_C , satisfy

$$I_E + I_B + I_C = 0, \quad (19-3)$$

while eq. 19-2 implies that I_B is small in magnitude (typically, of the order of tens of microamperes) compared to either of $|I_E|$ and $|I_C|$ (typically, of the order of several milliamperes).

19.4.5 The transistor in the saturation and cut-off modes

If the B-E diode of a transistor is forward biased (as in the active mode) while, at the same time, the C-B diode is also *forward biased* (in contrast to the active mode), the transistor is said to be in *saturation*. Here the C-B diode sends in a stream of electrons directed from the collector to the base, which largely cancels the current caused by the stream of electrons injected into the base from the emitter.

If, on the other hand, the B-E diode is *reverse biased*, the transistor is said to be in the *cut-off* mode. The B-C diode may be either in forward or in reverse bias, though the latter commonly turns out to be case. In this mode of operation, the stream of electrons injected from the emitter into the base is cut off, and all the three currents (I_E , I_C , and I_B) are reduced to nearly zero values.

19.4.6 Transistor characteristics

The three inter-electrode voltages (V_{BE} , V_{BC} and V_{CE}) for a transistor are not independent of one another, but satisfy, by definition,

$$V_{BE} - V_{BC} = V_{CE}. \quad (19-4)$$

Other inter-electrode voltages could also be defined. Instead of V_{BE} , for instance, one could refer to V_{EB} . However, one would have, $V_{EB} = -V_{BE}$.

Similarly, the three currents I_E , I_C and I_B satisfy among themselves the relation (19-3). Thus, one is left with two voltages and two currents characterizing the state of the transistor at any given instant of time. Even among these four quantities, there exist interrelations by virtue of the physical processes taking place in the transistor. It is found that *two* of these quantities can be made to vary independently, while two others get determined by these, depending on the characteristics of the transistor under consideration. The nature of such dependence can be depicted graphically. Such graphical relations are referred to as *transistor characteristics*.

One can, for instance, choose I_B and V_{CE} as the two independently variable quantities, and graphically express I_C , V_{BE} as functions of these (V_{BC} and I_E would then be determined from equations (19-4) and (19-3)). There exist, of course, other possible choices of independent variables and other ways of graphically expressing the current-voltage relations, but the one mentioned above happens to be a convenient way of expressing these relations. Moreover, of the two dependent variables chosen above, we consider below the graphical representation of the collector current I_C for the purpose of illustration since it gives a good idea of transistor operation that one can start from in designing and analyzing transistor circuits.

In expressing the dependence of I_C on V_{CE} and I_B graphically, I_B is commonly represented as a *parameter*, i.e., one plots I_C against V_{CE} for various fixed values of I_B , thereby obtaining a set of graphs. These are referred to as the *common emitter output*

characteristic curves (or *collector curves* in brief).

The idea underlying such nomenclature relates to a class of transistor circuits referred to as *common emitter* circuits. In the *common emitter amplifier* (see sec. 19.4.11), for instance, the emitter junction of the transistor is maintained at a constant potential (referred to as an *AC ground*) with no AC variation in it, and the potentials at the base and collector terminals are expressed with reference to the emitter potential. The transistor in such a circuit may be looked upon as a *two-port* device with the base- and the emitter terminals making up the *input* port, and the collector- and emitter terminals making up the *output* port. Notice that the emitter terminal here is *common* to the two ports. The instantaneous values of V_{BE} and I_B may then be looked upon as the input voltage and the input current respectively, while the instantaneous values of V_{CE} and I_C can similarly be looked upon as the output voltage and current respectively (more precisely, the *AC parts* of these voltages and currents are often of greater relevance in analyzing the functioning of the circuits). The common emitter output characteristics then relate the output current to the output voltage, with the input current used as a parameter.

Incidentally, when speaking of a common terminal, one has to keep in mind whether the voltages under consideration are DC or AC ones. For instance, an AC ground, i.e., a fixed value (0 V) of the AC voltage at a junction, need not always correspond to a DC ground as well.

Fig. 19-18 depicts schematically a typical set of collector curves for a n-p-n transistor, drawn for three fixed values (I_{B1} , I_{B2} , I_{B3} , with $I_{B1} < I_{B2} < I_{B3}$) of the input current I_B .

Referring to any one of the collector curves, one finds that it is made up of a sharply rising portion, followed by an approximately flat and linear variation of I_C with V_{CE} . The region containing the collector curves to the right of the dotted line in the figure corresponds to the transistor being in the *active* mode (see sec. 19.4.4), while the region lying to the left of the line corresponds to the *saturation* mode (see sec. 19.4.5). For a silicon transistor, the dotted line demarcating these two modes of transistor operation corresponds to $V_{CE} \approx 0.2$ V. Finally the region below the V_{CE} -axis corresponds to the

transistor being in the *cut-off* mode (the characteristic curves in the cut-off region are not shown in the figure).

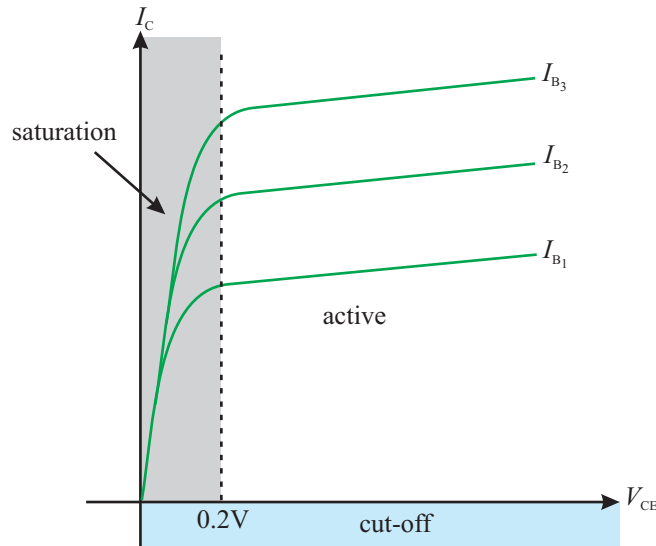


Figure 19-18: Common emitter output characteristic curves (schematic) for an n-p-n transistor; three curves are depicted, with three chosen values of the base current ($I_{B_1} < I_{B_2} < I_{B_3}$), where the collector current is plotted against the collector-emitter voltage for each base current; the active, saturation, and cut-off regions of transistor operation are indicated.

More precisely, the cut-off region is demarcated from the active and the saturation regions not by the line $I_C = 0$, but by the line $I_E = 0$. For many practical purposes, however, this distinction can be ignored.

It is not difficult to explain the nature of the transistor characteristic curves in these three regions, at least in a qualitative manner. Starting with the active region, in which the B-E diode is forward biased ($V_{BE} \approx 0.6\text{-}0.7$ V for a silicon transistor; a value $V_{BE} = 0.7$ V is commonly assumed) while the C-B diode (or the B-C diode, whichever one prefers) is reverse biased ($V_{CB} > 0$), one observes that V_{CB} is positive in this region where, moreover, the variations in V_{CE} are almost entirely due to those in V_{CB} . The latter has little effect on I_C since the role of the collector here is simply to collect most of the electrons injected from the emitter into the base region. For any given value of I_B , the current due to these electrons is, to a first degree of approximation, a constant (determined by the constant

β of the transistor, see sec. 19.4.7), and one thus finds little variation of $I_C (\approx -I_E)$ with V_{CE} in this region.

Looking more closely at this portion of a collector curve, one observes, however, a slow rise of the collector current with V_{CE} . This is explained by a small effect exerted by the C-B diode on the B-E diode, which I prefer not to enter into in this introductory exposition.

As V_{CE} is made to decrease to comparatively low values, a stage is reached when the C-B diode becomes forward biased. In order for the forward bias to be effective one has to have $V_{CB} \approx -0.7V$ (recall that we are referring to a n-p-n transistor), i.e., $V_{CE} \approx 0V$. As the C-B diode becomes forward biased, it sends in a stream of electrons from the collector to the base region which cancels to a large extent the current due to the electrons injected from the emitter into the base, as a result of which the collector current drops sharply. This explains the sharp bend downward of the collector curves in the saturation region. For a silicon transistor the process of reduction of the collector current starts from $V_{CE} \approx 0.2V$, i.e., as V_{BC} crosses the level $\approx 0.5V$.

Finally, as the base-emitter diode is made reverse biased, the injection of the stream of electrons from the emitter into the base region ceases and one has, as a consequence, $I_C \approx 0$. This explains the cut-off region of operation below the V_{CE} -axis in the transistor characteristics.

In addition to the common emitter output characteristics, one also has to refer to the common emitter *input* characteristics for the sake of completeness, where one plots V_{BE} as a function of I_B for various fixed values of V_{CE} . These curves, drawn for various different values of V_{CE} , are closely bunched together, and resemble the diode characteristic curve (relating to the B-E diode) of fig. 19-10. However, these plots will not be of direct relevance for our present purpose.

19.4.7 The parameters α and β of the transistor

Continuing to refer to a n-p-n transistor one observes that, the way the transistor currents and voltages are defined, the collector current I_C is positive in the active region and the emitter current I_E is negative, while the magnitudes of the two currents are almost equal ($I_C \approx -I_E$).

Since the emitter injects a stream of electrons into the base region, there occurs a corresponding flow of electrons from the external circuit *into* the emitter region. On the other hand, the stream of electrons entering into the collector *leaves* the collector junction into the external circuit.

A more accurate representation of the relation between the emitter- and the collector currents in the active region of the transistor characteristics is

$$I_C = -\alpha I_E + I_{C0}. \quad (19-5)$$

This is an improvement over the approximate relation $I_C = -I_E$ on two counts. First, it takes into account the fact that not all the electrons injected from the emitter into the base succeed in finally reaching the collector, since a few of these electrons recombine with holes in the base region. Supposing that a fraction α of the electrons entering into the base region cross it over to the base-collector junction, the magnitude of I_C will be reduced from that of I_E by the same fraction, which explains the first term of the relation (19-5). In this relation, α a characteristic parameter of the transistor whose value is less than but close to unity. It does depend to a small extent on the values of I_B and V_{CE} , but for our present purpose we can take it to be a constant characterizing the transistor.

The relation (19-5), moreover, includes a term I_{C0} , which is the reverse current flowing through the B-C junction, the direction of this current being the same as that of the total collector current I_C . Though it constitutes only a small fraction of I_C , it is at times necessary to take it into account in the analysis of transistor circuits, especially because

of its pronounced temperature dependence.

One can combine the relation (19-5) with (19-3) to work out a relation between I_B and I_C in the active region:

$$I_C = \beta I_B + (1 + \beta)I_{C0}, \quad (19-6a)$$

where

$$\beta = \frac{\alpha}{1 - \alpha}. \quad (19-6b)$$

The parameter β is once again a characteristic of the transistor under consideration, for most purposes a more useful one than α . Its value is ordinarily of the order of 100. Thus a transistor with β as low as, say, 10, or as high as 500 is likely to be a defective one. Since, in the active region, I_{C0} is usually small compared to I_B , the relation (19-6a) is often approximated by the simpler form

$$I_C \approx \beta I_B. \quad (19-7)$$

The parameter β defined as above with reference to DC values of the transistor currents is sometimes referred to as the *DC beta* of the transistor. A more commonly used parameter is the *AC beta* defined as

$$\beta_{AC} = \frac{\delta I_C}{\delta I_B}, \quad (19-8)$$

where δI_C and δI_B stand for *small variations* in I_C and I_B respectively (where, technically speaking, V_{CE} is to be assumed to be held constant; however, this constraint is usually not of much practical importance). Often, one uses the symbol h_{fe} (the *common emitter forward current gain*) instead of β_{AC} (which, for practical purposes, can be taken to be the same as β_{DC}), where h_{fe} is one of *four* basic parameters characterizing a transistor, collectively called the *common emitter hybrid parameters*. Of these four, two are commonly not of much relevance in the analysis of transistor circuits while, of the remaining two, one is the parameter h_{fe} introduced above. The remaining parameter, denoted h_{ie} I

will briefly introduce below.

The parameter β as also the parameters h_{fe} and h_{ie} depend on the state of the transistor specified by the values of I_C and V_{CE} (or any other appropriate pair of variables; the values of these two define what is termed the *Q-point* of the transistor, see sec. 19.4.11.1 below). Among these, the variation of h_{ie} is the most pronounced. In other words, h_{ie} is dependent on the *Q*-point to a greater extent than h_{fe} , where it is assumed that the *Q*-point lies in the active region of the transistor characteristic curves.

19.4.8 Convention for using notations

A few words on *notation* are in order here. The symbols like I_C and V_{CE} are used to denote *DC* currents and voltages in the transistor, i.e., ones that do not vary with time. The currents and voltages in most transistor circuits are, however, *alternating ones*, with values changing with time. Symbols such as i_C and v_{CE} are used to denote the *instantaneous* values of these quantities at any given point of time. Thus, the *mean* or average of these instantaneous values taken over a cycle of the AC gives the DC quantities like I_C and V_{CE} . Finally, the *AC parts* of the currents and voltages are denoted by symbols such as i_c and v_{ce} which means that, for instance, $i_c = i_C - I_C$. At times, these AC parts are represented as small variations in the instantaneous values or small variations that may be brought about in the DC values so that, for instance, $i_c = \delta I_C$. Here δI_C denotes a small change in the DC collector current that may be introduced in a circuit by making small changes in the circuit parameters.

19.4.9 The common emitter input impedance

The common emitter input impedance can now be defined as

$$h_{ie} = \frac{\delta V_{BE}}{\delta I_B} = \frac{v_{be}}{i_b}, \quad (19-9)$$

where the notations introduced in sec. 19.4.8 have been made use of. However, in order to make the definition complete one has to add the condition that the AC base-emitter

voltage and the AC base current are to be evaluated under the constraint of a constant collector-emitter voltage, i.e., $\delta V_{CE} = v_{ce} = 0$.

This definition can be interpreted as saying that h_{ie} is the AC input resistance (a more commonly used term in the context of AC circuits is, input *impedance*) of the transistor in the common emitter mode of description, i.e., the effective resistance that an ideal AC voltage source would face when connected between the base and the emitter.

A related quantity of considerable importance is the AC resistance of the base-emitter diode, which is defined as

$$r_e = -\frac{v_{be}}{i_e}, \quad (19-10)$$

where the negative sign comes from the defined direction of i_e (similar to I_E) and where, for the sake of completeness, one has to add the constraint of a constant value of V_{CB} .

The two parameters h_{ie} and r_e are related as

$$h_{ie} \approx \beta r_e \quad (19-11)$$

where the symbol β is often used in the place of h_{fe} .

The value of h_{ie} for a transistor ranges between several hundred ohm to a few kilo-ohm, depending on I_C and V_{CE} , while that of r_e is of the order of a few ohm.

19.4.10 AC transistor operation: summary

In summary, the transistor in the active mode in the common emitter configuration presents a resistance of several hundred ohm to a source of AC voltage connected to its input port (between the base and the emitter terminals) and multiplies the input AC current (i_b) by a factor of h_{fe} (≈ 100) to produce an output current i_c . Looked at from the output terminals (between the collector and the emitter) as a current source, the transistor has a high output impedance ($\frac{\delta v_{CE}}{\delta I_C}$, $\delta I_{r_{mC}}$ small). Such a high output

impedance for a current source can be altogether ignored since the output impedance operates in parallel with the load resistance to which the current is supplied.

19.4.11 Voltage amplification

Fig. 19-19 depicts the basic circuit of a *common emitter transistor amplifier*, used in voltage amplification. To the left of the figure is an AC voltage source supplying AC voltage to the circuit to its right. The latter in turn supplies an amplified AC voltage to the load resistance R

1. In an actual application, the amplifier may supply the amplified voltage to *another* amplifier stage or to some other circuit module. In that case, R would stand for the *input impedance* of the next stage. Moreover, the voltage source from which the amplifier receives its input AC voltage may itself be a voltage amplifier or an earlier stage of a larger circuit.
2. The circuit shown in the figure is seldom used in the construction of an actual amplifier. It is given here so as to illustrate the basic principles of voltage amplification with a transistor.

19.4.11.1 DC bias: the Q-point

In the circuit of fig. 19-19, the resistances R_B , R_E , and R_C , and the DC voltage sources of EMF V_{BB} and V_{CC} are employed so as to realize a set of desired values of the DC currents and voltages (such as I_C and V_{CE} ; in actual operation these are the averages of the corresponding instantaneous currents and voltages) or a desired *Q-point* of operation for the circuit. These DC quantities are referred to as *bias* currents and voltages.

By employing standard techniques of DC circuit analysis along with the graphical representation of the transistor characteristics, one can work out the values of the DC currents and voltages such as I_C and V_{CE} for any given set of values of the circuit parameters R_B , R_E , R_C , V_{BB} , and V_{CC} . These DC currents and voltages define the *Q-point* in the graph depicting the transistor characteristics (see fig. 19-20).

The *AC operation* of the transistor (such as the amplification of an applied AC voltage in

the present context) depends on the location of the Q -point within the set of characteristic curves. For instance, if the Q -point is located within the saturation or the cut-off regions, the desired AC operation will not be realized. Put differently, in order to make a transistor perform its desired AC operation appropriately, one has to arrange for an appropriate set of values of the bias currents and voltages by employing appropriate circuit parameters in the biasing circuit (the part of the circuit including R_B , R_E , and R_C , V_{BB} , and V_{CC} in the present instance). These DC bias currents and voltages correspond to a desired Q -point within the set of characteristic curves of the transistor.

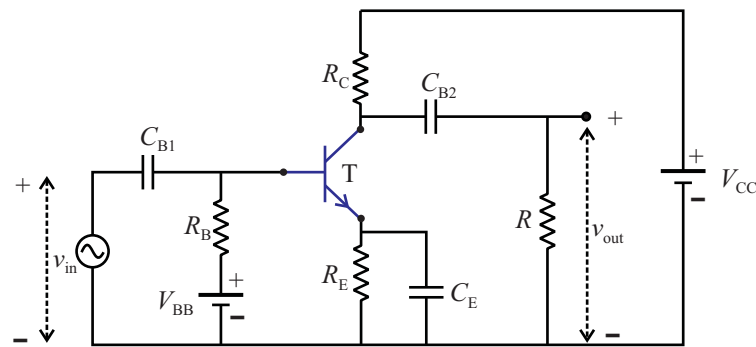


Figure 19-19: The basic amplifier circuit using an n-p-n transistor; DC voltage sources with voltages V_{BB} and V_{CC} and the resistances R_B , R_C , R_E are used to bias the transistor, setting up desired values of I_C and I_B ; the AC source causes a periodic variation in I_B and I_C , and an AC voltage drop across the resistance R ; the three capacitors (with capacitance C_{B1} , C_{B2} and C_E) are used so as to enable the set-up to perform the amplification of the input voltage (v_{in}) supplied by the AC source, the output voltage (v_{out}) being that appearing across R .

The Q -point chosen appropriately within the set of characteristic curves has, moreover, to be a *stable* one, i.e., its location within the characteristic curves must not change appreciably for small but unavoidable changes in the circuit conditions (such as the circuit parameters mentioned above or the ambient temperature). This requires additional considerations in the designing of the biasing circuit.

The reason why the Q -point is important is that the parameters (such as h_{ie} and h_{fe}) characterizing the AC operation of the transistor depend on the location of the Q -point. Considering the active region of the characteristic curves, the variation of h_{ie} with the

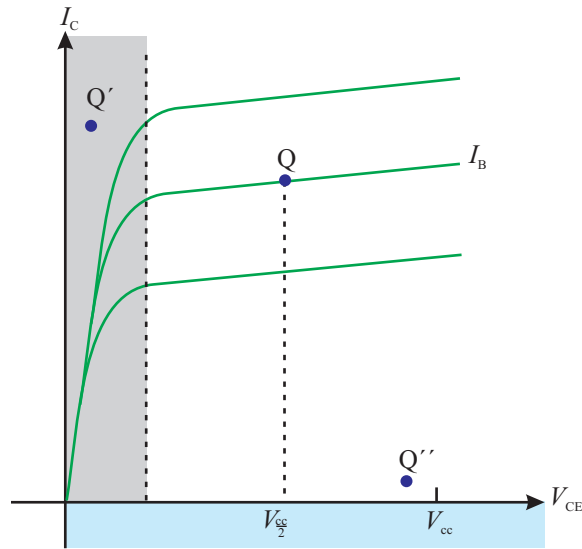


Figure 19-20: Illustrating Q-point and its stability; a Q-point with $V_{CE} \approx \frac{V_{CC}}{2}$ is generally considered to be a good choice; in addition, an appropriate value of I_B is to be chosen; the Q-point is to be stable, i.e., it must not change its location appreciably within the set of characteristic curves due to unavoidable changes, if any, in the circuit parameters; a Q-point located at, say, Q' or Q'' is not appropriate for a voltage amplifier.

location of the Q-point is more pronounced than that of h_{fe} . A drastic alteration of the values of the transistor parameters takes place when the Q-point moves over from the active region to the saturation or the cut-off region. In the case of the common emitter amplifier, the Q-point is chosen somewhere in the middle of the active region of the characteristic curves, the condition for which is commonly expressed as $V_{CE} \simeq \frac{V_{CC}}{2}$.

19.4.11.2 Blocking and bypass capacitors

In addition to the circuit elements for realizing the appropriate Q-point and for achieving its stability, the circuit of fig. 19-19 includes the capacitances C_{B1} , C_{B2} and C_E . Of these, C_{B1} is used so as to prevent DC voltages, if any, coming from the voltage source (or from an earlier stage of a larger circuit), from being transferred to the base-emitter terminal of the transistor and upsetting the Q-point. Similarly, the capacitance C_{B2} prevents the DC voltage, if any, generated by the amplifier, from getting transferred to the load resistance R or to succeeding stages of a larger circuit. This is why the two capacitors corresponding to C_{B1} and C_{B2} are referred to as *blocking capacitors*.

The capacitance C_E , on the other hand, serves the purpose of *bypassing* the resistance R_E , where the latter is required for achieving bias stability. Unless bypassed, this resistance causes a *negative feedback* to appear in the AC operation of the amplifier, thereby considerably reducing its *voltage gain* (see sec. 19.4.11.3).

The capacitance C_E is so chosen as to make its impedance (see sec. 13.5.3) small in magnitude at the operating frequency, as a result of which the AC voltage of the emitter remains zero (i.e., in other words, there is little variation of the emitter voltage away from the common or ground potential). If the capacitance were not used, then there would be a considerable variation in the emitter voltage due to the AC voltage appearing across R_E which would *cancel* a large part of the input voltage to the amplifier. This is referred to as *negative feedback* in an electronic circuit. The term feedback refers to an effective coupling between the output and input stages of a circuit.

Though C_E is used to remove the negative feedback that would possibly be caused by R_E , the presence of negative feedback to a certain degree may nevertheless prove to be useful in the performance of an amplifier. This, however, calls for more elaborate considerations in the designing of the amplifier.

19.4.11.3 AC operation of the amplifier: voltage gain

With all this background, it is now time to have a look at how the voltage amplification takes place.

Reduced to bare essentials, the voltage amplification in the circuit of fig. 19-19 consists of an amplification of the AC base-emitter voltage (v_{be}), the *input* voltage of the amplifier, resulting in the *output* voltage, which in the present context is the AC collector-emitter voltage (v_{ce}).

This leads one to define the *voltage gain* of the amplifier as

$$A_v = \frac{v_{ce}}{v_{be}}, \quad (19-12)$$

where, moreover, one assumes an infinitely large load resistance ($R \rightarrow \infty$) in working

out the above ratio.

On the face of it, the condition $R \rightarrow \infty$ may appear puzzling since it is the load resistance R across which the output voltage appears. Looking more closely, one realizes that, since the emitter is effectively at AC ground because of the bypassing action of C_E , one end of R is connected to the collector and the other end effectively to the emitter, as long as AC voltages at these are under consideration. Similarly, since the positive terminal of the DC voltage source V_{CC} is at a fixed voltage, the terminals of R_C are also connected effectively between the collector and the AC ground, i.e., the emitter. In other words, R_C and R are connected in parallel between the collector and the emitter in the context of the AC operation of the circuit. The AC output voltage thus appears across this parallel combination. The assumption of $R \rightarrow \infty$ (load *open-circuited*) merely removes the dependence of the voltage gain on R , making it a characteristic of the amplifier circuit alone. Indeed, R does not *belong* to the amplifier circuit, it is merely the external device to which the amplifier delivers the amplified voltage.

In other words, the voltage gain is the ratio of the AC voltage across R_C and that between the base and the emitter. The latter voltage is given, to a good degree of approximation, by product of the AC diode current and the AC diode resistance of the base-emitter diode, i.e.,

$$v_{be} \approx -i_e r_e \approx i_c r_e, \quad (19-13)$$

where, in the last expression we have made use of the approximate relation $i_c \approx -i_e$.

The AC output voltage is, on the other hand, produced by the passage of the current i_c through R_C :

$$v_{ce} = -i_c R_C, \quad (19-14)$$

where the negative sign arises due to the assumed direction of i_c and the assumed

polarity of v_{ce} (check this out).

One therefore ends up with the following expression for the voltage gain of the amplifier,

$$A_v = -\frac{R_C}{r_e}. \quad (19-15)$$

The basic idea underlying voltage amplification can now be formulated. Essentially the same AC current (i_c ; recall that $i_e \approx -i_c$) flows through the low diode resistance r_e (a few ohms) of the forward biased base-emitter diode and through R_C , which is typically of the order of a kilo-ohm. The potential drop across these two correspond to the input and to the output voltages of the circuit. This results in a large value of the ratio of the latter to the former, explaining the voltage gain.

R_C cannot be increased arbitrarily so as to achieve a high voltage gain because too large a value of R_C leads to *distortion* in the output voltage.

The negative sign in eq. (19-15) indicates that, along with voltage amplification, the transistor reverses the polarity of the input voltage in producing the output voltage.

Voltage amplification can be explained from another, equivalent, point of view as well, by noting that the AC input voltage can also be expressed as the potential drop due to the input current i_b flowing through the input resistance h_{ie} of the amplifier, i.e., $v_{be} = h_{ie}i_b$. This leads to the same expression for the voltage gain as that in eq. (19-15) when one makes use of the relations (19-11) and (19-8) (recall that the symbol β is often used by default to denote β_{AC} , i.e., h_{fe}).

The entire analysis of the common emitter amplifier given above applies provided that the frequency of the input voltage is neither too low nor too high. Indeed, for the purpose of analysis, one has to consider three different ranges of frequencies, namely, the *low* frequency, the *mid*-frequency, and the *high* frequency ranges. The analysis given above applies to the mid-frequency range, from a few hundred Hz to frequencies of the order of 100 kHz. In this mid-frequency range the voltage gain given by eq. (19-15) is essentially

independent of the frequency. On either side of this range, however, i.e., for lower and higher frequencies, the voltage gain *drops* to low values. An explanation of these frequency effects needs more detailed considerations which I prefer not to go into in this book.

Problem 19-5

In the transistor amplifier circuit of fig. 19-19 the DC base- and collector currents are $50\ \mu\text{A}$ and $5.00\ \text{mA}$ respectively, where the transistor operates in the active mode. What is the value of the parameters β and α under this operating condition? If the input impedance of the transistor is $h_{ie} = 500\ \Omega$, and the collector resistance used is $R_C = 600\ \Omega$, find the voltage amplification factor A_v .

Answer to Problem 19-5

HINT: The parameter β is given by $\beta = \frac{I_C}{I_B} = \frac{5 \times 10^{-3}}{50 \times 10^{-6}} = 100$, where the transistor is assumed to be operative in the active mode. Hence the parameter α is $\alpha = \frac{\beta}{1+\beta} = 0.99$ (approx; see eq. (19-6b)). One can assume that the AC value of the transistor current gain ($h_{fe} = \frac{\delta I_C}{\delta I_B}$) is also 100. Since the input impedance h_{ie} is $500\ \Omega$, the base-emitter AC diode resistance is $r_e = \frac{500}{100}\ \Omega = 5\ \Omega$. Hence the magnitude of the voltage gain is $|A_v| = \frac{R_C}{r_e} = 120$.

19.5 The operational amplifier (Op-amp)

The *operational amplifier* ('op-amp' in brief) is a voltage amplifier with near-ideal performance features that is almost universally employed in electronic circuits where it serves a variety of purposes, all resulting from its basic function as voltage amplifier of unique efficacy. It is a small scale integrated circuit, marketed as a small eight-terminal ('eight-pin') 'chip'.

Basically, an op-amp is a *differential amplifier*, followed by gain-enhancing and buffer stages, the latter designed so as to set the input- and output impedance of the device at appropriate values (see below).

19.5.1 The differential amplifier

A differential amplifier, which essentially consists of two amplifiers with equal and opposite gains ($A_1, A_2; A_1 \approx -A_2$), is schematically represented in fig. 19-21, in which A and B are input terminals and C is the output terminal. Voltages v_1, v_2 (with reference to the common terminal G, referred to as the ‘ground’; we refer to AC voltages for the sake of concreteness; however, the op-amp is a ‘direct-coupled’ amplifier where DC voltages can also be used as inputs) are presented at the terminals A, B, when the voltage appearing at the output terminal C is

$$v_o = A_1 v_1 + A_2 v_2 = A_1 \left(v_c + \frac{v_d}{2} \right) + A_2 \left(v_c - \frac{v_d}{2} \right), \quad (19-16a)$$

where

$$v_c \equiv \frac{v_1 + v_2}{2}, \quad v_d \equiv v_1 - v_2, \quad (19-16b)$$

are referred to, respectively, as the *common mode* and the *difference* input voltage to the amplifier. Making use of the relation $A_1 \approx -A_2$, and defining the *differential gain* as

$$A_d = \frac{A_1 - A_2}{2}, \quad (19-16c)$$

one can express the output voltage as

$$v_o \approx A_d v_d, \quad (19-16d)$$

which holds to a good degree of approximation provided that the *common mode gain*

$$A_c = A_1 + A_2, \quad (19-16e)$$

is small, so that the *common mode rejection ratio* (CMRR)

$$\rho = \frac{A_d}{A_c}, \quad (19-16f)$$

has a high value.

A differential amplifier can be set up with two identical common emitter amplifiers with their emitter terminals joined together.

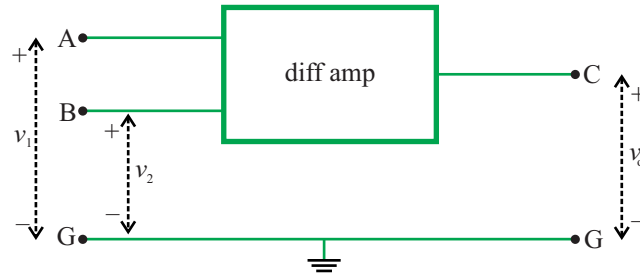


Figure 19-21: Schematic representation of a differential amplifier; A and B are the two input terminals where voltages v_1, v_2 (both with reference to the ground G) are applied; C is the output terminal where the output v_o , which is an amplified difference voltage $v_d (=v_1 - v_2)$, is produced; the '+' and '-' signs are used to indicate the definitions of the respective voltages.

Problem 19-6

When inputs $v_1 = 0.5 \text{ mV}$ and $v_2 = -0.5 \text{ mV}$ are applied to the terminals of a differential amplifier, an output of 0.1 V is obtained, while an input of 0.5 mV to both the terminals results in an output of 0.1 mV . Calculate the CMRR and obtain the output for input voltages $v_1 = 0.6 \text{ mV}$, $v_2 = 0.3 \text{ mV}$.

Answer to Problem 19-6

SOLUTION: With $v_1 = v_2 = 0.5 \text{ V}$, the difference voltage is $v_d = 0$, and one obtains, from (19-16a), $v_o = A_c v_c$. Using given values, the common mode gain is found to be $A_c = 0.2$. On the other hand, with $v_1 = 0.5 \text{ mV}$ and $v_2 = -0.5 \text{ mV}$, and making the approximation $A_c \approx 0$ for the sake of convenience, one obtains, from (19-16d), $A_d = \frac{v_o}{v_d} = \frac{0.1}{10^{-3}} = 100$. Hence the CMRR is $\rho = \frac{100}{0.2} = 500$. On making the same approximation with $v_1 = 0.6 \text{ mV}$, $v_2 = 0.3 \text{ mV}$, the output voltage is found to be $v_o = A_d v_d = 100 \times 0.3 \text{ mV}$, i.e., 30 mV .

19.5.2 Op-amp basics

The differential amplifier makes up the input stage of the operational amplifier and is followed by amplifier stages enhancing the overall gain. Other components are included

in the circuitry to enhance the *input impedance* and to reduce the *output impedance* of the set-up.

The concept of input and output impedance of an electronic device is an important one. When a voltage source is connected to the input terminals of the device, the effective impedance presented to the source is termed its input impedance. The output impedance is similarly defined with reference to the output terminals of the device.

The op-amp is a *direct-coupled* device, which means that there are no reactive elements blocking the DC components of the various signals at the input terminals, in between stages, and at the output terminals. In the case of voltage signals, for instance, the DC components are prevented from being transferred from one stage to another by means of *blocking capacitors* (recall the use of such capacitors in a common emitter amplifier). The use of such capacitors is not convenient in integrated circuits which is why the op-amp is direct-coupled (or *DC-coupled*). As a result, it is capable of amplifying DC as well as AC voltages.

Stated briefly, an op-amp is a DC-coupled voltage amplifier that has a very high voltage gain (A_v), a very high input impedance (Z_{in}), and a very low output impedance (Z_{out}), all of which are desirable performance characteristics for a voltage amplifier. The fact that Z_{in} is high ($\rightarrow \infty$) means that when an op-amp receives a voltage signal from a source, the voltage received will differ little from the open circuit voltage of the source, regardless of the output impedance of the latter. Similarly, a very low value of Z_{out} (≈ 0) means that the voltage delivered to a load resistance will be independent of the value of the latter.

The high gain ($A_v \rightarrow \infty$) is a special feature of the op-amp and is responsible for many of its applications, but at the same time tends to make its operation unstable. For instance, even very small noise signals may be picked up and amplified. Moreover, the gain itself may vary over wide ranges due to, say, temperature fluctuations. In practice, however, the gain is moderated by employing appropriate *negative feedback* (i.e., an arrangement where a part of the output voltage is mixed with the input voltage, with a

phase opposite to that of the latter). Thus, A_v is termed the *nominal gain* of the op-amp while the *actual gain* differs markedly from this and may even be quite small, depending on the purpose at hand. What is more, this actual gain can be made very stable, being determined by circuit parameters that can be accurately controlled.

In addition to the above features, the op-amp possesses other desirable characteristics as well, namely, for instance, a large *bandwidth*. The term 'bandwidth' signifies the range of frequencies over which the gain (along with the input impedance, and the output impedance) of an amplifier remains constant (or nearly so), varying only outside this range. With appropriately employed negative feedback, the bandwidth of an op-amp can indeed be made very large.

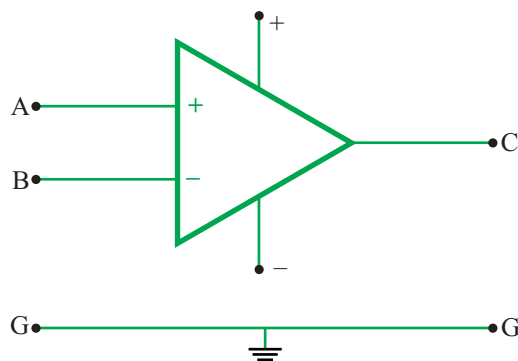


Figure 19-22: The circuit symbol of an operational amplifier (op-amp); A and B are, respectively, the *non-inverting* and the *inverting* input terminals to which two input voltages (with reference to a common or 'ground' terminal G) are presented, the output voltage (at terminal C, again with reference to the ground) being the amplified version of the difference of the two; the op-amp needs two DC bias voltages, fed at terminals marked '+' and '-' (to be distinguished from the '+' and '-' marks inside the triangular symbol, meant to designate the non-inverting and the inverting input terminals).

Fig. 19-22 depicts the circuit symbol of an op-amp. A and B are, respectively, the *non-inverting* and the *inverting* input terminals to which two input voltages (with reference to a common or 'ground' terminal G) are presented, the output voltage (at terminal C, again with reference to the ground) being the amplified version of the difference of the two. The op-amp needs two DC bias voltages, fed at terminals marked '+' and '-'. The op-amp IC 741 (a commonly available chip) has two other terminals available, of which

one is termed the *offset null* (used to adjust the output to zero level when both the input terminals are grounded), while the other has no connection to it.

Two common applications of the op-amp are in the form of the *inverting* and *non-inverting* amplifiers. In the former, the output voltage appears with the same phase as the input voltage while, in the latter, the two phases are opposite to each other. In both these amplifiers, negative feedback is employed so as to achieve stability of operation, and the effective gain is controlled by circuit parameters external to the op-amp (recall that the gain of the op-amp itself is very high, engendering an inherent instability of operation).

19.6 Oscillators

Amplifiers are commonly provided with a negative feedback so as to make their operation stable against fluctuations of various kinds. Conversely, a positive feedback (where a part of the output voltage is mixed with the input voltage at such a phase difference as to effectively add to the latter) has the effect of making the amplifier operation unstable. This is made use of in setting up an *oscillator* where the instability is made to reach a critical level such that the set-up delivers a steady oscillating output even *without* an input EMF from an AC source. What the oscillator does is that it amplifies the fluctuations inherent in the DC voltage sources (supplying the bias voltages to the set-up), picking out some specific AC component of the fluctuations and amplifying it in a self-sustaining manner.

Fig. 19-23 depicts the set-up of a *Wien bridge oscillator* in which an op-amp is made use of, and a positive feedback is employed, with the op-amp used as a non-inverting amplifier.

I want you to get the picture clear as to what is happening at which level. Firstly, the op-amp itself is an amplifier with a very high gain. External circuitry is added to the op-amp so as to reduce the effective gain to a stable value, making the op-amp operate as a non-inverting amplifier. At the next level of circuitry, positive feedback

is employed to this non-inverting amplifier (i.e., the op-amp along with the negative feedback mechanism; however, no AC input is given to this amplifier stage) so as to make it operate as an oscillator.

The arrangement for the positive feedback (termed a 'lead-lag loop') is shown to the left of the dashed lined in the figure, while the op-amp along with the negative feedback mechanism (making up a non-inverting amplifier) is shown to the right of the same line. Taking a hard look at the circuit, one can discover a similarity with the circuit of fig. 13-25, the labels A, B, C, D indicating corresponding junctions in the two circuits. In the present instance, the AC voltage from the source in fig. 13-25 is replaced with the output voltage of the op-amp, while the input terminals of the latter are connected to the junctions B, D instead of the terminals of the AC ammeter. The four impedances $\tilde{Z}_1, \tilde{Z}_2, \tilde{Z}_3, \tilde{Z}_4$ of the AC Wheatstone bridge are now to be taken as $\tilde{Z}_1 = R + \frac{1}{j\omega C}$, $\tilde{Z}_2 = \frac{1}{\frac{1}{R} + j\omega C}$, $\tilde{Z}_3 = R_1$, $\tilde{Z}_4 = R_2$, where R_1, R_2, R, C are as shown in the figure, and ω is the frequency at which the circuit operates, as mentioned below.

The balanced condition of the bridge corresponds to zero input to the op-amp which produces a steady output at a frequency ω (it is this frequency which appears in the expressions for \tilde{Z}_1, \tilde{Z}_2 given above) because of its effectively infinite gain. On making use of the relation (13-71) one obtains the condition for steady oscillation produced by the arrangement and the frequency of the output voltage as

$$R_1 = 2R_2, \omega = \frac{1}{CR}. \quad (19-17)$$

19.7 Introduction to digital electronics

Digital electronics involves circuits in which there are only two possible states of relevance at any point in a circuit at any given instant of time, the states being represented by *voltage levels*. Alternative representations in terms of current levels, glow and extinction of lamps, etc., are also possible, but what is common to all these modes is the *discreteness* of the states based on which the circuits function.

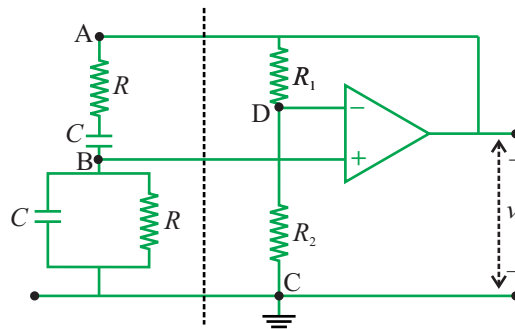


Figure 19-23: The circuit diagram for a Wien bridge oscillator employing an op-amp, with a lead-lag loop to the left of the dotted line and the non-inverting amplifier to the right of the same line; R_1, R_2 are feedback resistances making up the non-inverting amplifier, satisfying the first relation in (19-17), while R, C are circuit parameters making up the lead-lag loop, in terms of which the frequency of the oscillations produced by the circuit is given by the second relation in (19-17); v_o is the AC output voltage; the bias terminals of the op-amp are not shown, while the '-' and '+' signs within the op-amp symbol denote the inverting and the non-inverting terminals of the op-amp.

Digital electronics made its appearance in the field of applied science in the late forties of the last century when the computer was first introduced. Since then, the subject has progressed in enormous strides, making possible the wonders of modern computer technology. Moreover, digital electronics is now no longer confined to computer technology alone. It is ubiquitous in the fields of communications, instrumentation, control, and in numerous other areas permeating the whole of modern technology.

19.7.1 Boolean algebra

When represented in terms of voltage levels, the two possible states at any point in a digital circuit can be labeled as *high* (H) and *low* (L) respectively. For instance, any voltage in the range 0 V to 1.4 V may be labeled as *low* while any voltage in the range 3.6 V to 5.0 V may be labeled as *high*, while the circuit may be so designed that other values of voltage are excluded. What is important here is that the categorization of a voltage as *high* or *low*, depends only on the *range* in which it lies rather than on the actual value of the voltage in this range. Thus, a voltage measuring 1.1 V and one measuring 0.8 V will *both* be labeled as L in the circuit under consideration, and will both lead to similar functioning of the circuit.

The inter-relation of states at various points in the circuit, all described in the above manner can be expressed as operations or relations of *Boolean logic* or *Boolean algebra*.

From a mathematical point of view, the terms 'Boolean logic' or 'Boolean algebra' have specific meanings. However, we will use the terms loosely in a sense that will become apparent from the context.

In the Boolean algebra relevant in electronics, for instance, there are two basic Boolean 'numbers' 0 and 1 that can be made to correspond to *L* and *H* voltage states referred to above. These 'numbers' are the two *constants* in Boolean algebra, apart from which one can also refer to *variables* that can assume either of these two values. The theory further makes use of three *operations*, and functions or expressions made up of these constants, variables, and operations.

For instance, an operation of Boolean algebra is referred to as OR and is commonly denoted by the symbol '+'. If, then, *A* stands for a Boolean variable, then $A + 1$ is a Boolean expression made up of the variable *A*, the constant 1, and the operation '+'.

All this is related to the precise mathematical meaning of the term 'Boolean algebra', which we need not enter into here.

19.7.1.1 Digital circuits and binary numbers

Apart from the above interpretation of the state at any terminal or point in a digital circuit in terms of the elements of a Boolean algebra, the collective state at a number of terminals in the circuit can be made to correspond to *binary* representation of *numbers*, or representations (e.g., *octal*, *hexadecimal*) derived from binary ones. The voltage states *L* and *H* then correspond to the *binary digits* 0 and 1. Such a binary digit is called a *bit*. The binary system of representations of numbers is based on the two bits, 0 and 1. For instance, the number 4 in the commonly used decimal system of representation, assumes the form 100 in the binary system. By making the bits 0 and 1 corresponds to the states *L* and *H* respectively in a digital circuit, one can establish a correspon-

dence between the *collective states* at various points of the circuit, expressed in terms of Boolean algebra, and *numbers* in the binary representation.

Thus, there are *three* levels of interpretation of the states at various points in a digital circuit: (i) as electrical quantities of various magnitudes characterized in a discrete manner, (ii) as Boolean numbers or variables related through the operations of Boolean algebra, or (iii) as binary digits.

19.7.1.2 Combinational and sequential circuits

A digital circuit commonly involves a number of terminals whose states (*high* or *low*) are controlled externally, depending on the purpose and choice of one who operates the circuit. These are referred to as the *input* terminals of the circuit. The internal functioning of the circuit then determines the states at a number of *other* terminals, termed the *output* terminals of the circuit, consistent with the purpose the circuit is needed to serve. The relation between the output and the input states can be expressed in terms of Boolean operations.

Digital circuits, or components thereof, can be broadly classified into two categories, depending on the way they function. Some basic digital circuits produce outputs that are uniquely determined by the current values of the inputs. The inter-relation between the inputs and the outputs of such circuits can be expressed as specific combinations of Boolean operations. These are known as *combinational* circuits, and their functioning is said to follow combinational logic. The basic building blocks of such circuits are *logic gates*. The three elementary gates are the NOT, OR, and the AND gates, corresponding to three basic Boolean operations. Two other gates, namely the NOR and the NAND gates enjoy a wider applicability.

In the other category of digital circuits, the outputs depend not only on the current values of the inputs, but on the *past* sequence of states of the circuit as well. These are termed *sequential* circuits, and their functioning is said to follow sequential logic. The basic ingredients of sequential circuits are *flip-flops*, which are elementary digital memory units.

I am now going to relate to you a number of basic features of logic gates, and of binary numbers and binary arithmetic, and will then conclude this chapter with a brief introduction to flip-flops.

19.7.2 The basic logic gates

A logic *gate* is a piece of circuitry in which the output - commonly, one or several voltages - is related to the input (once again, one or more voltages) by means of some basic Boolean operation or other. Fig. 19-24 (A, B, C) gives the circuit symbols of the NOT, OR, and AND gates - the three basic or elementary gates of digital electronics. Associated with each of these logic gates (or with any other digital circuit) is a *truth table* giving in tabular form the relation between the input and the output variables of the circuit.

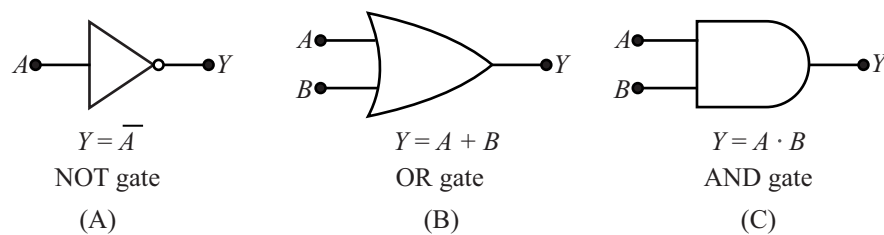


Figure 19-24: Circuit symbols for the (A) NOT, (B) OR, and (C) AND gates; in (A) A denotes the single input, while in each of (B) and (C), A and B denote the two inputs (OR and AND gates with more than two inputs are also possible); for each of the gates Y denotes the output variable; the relation between the input and the output Boolean variables can be expressed either in the form of a Boolean equation as shown in this figure below the circuit symbol of each of the gates, or in the form of a truth table.

The OR and AND gates shown in the figure are two-input ones, while OR and AND gates with more than two inputs are also used. The NOT gate, on the other hand has a single input.

The truth tables shown below are to be read as follows. For the NOT gate, for instance, A and Y stand for the values of the input and the output Boolean variables which, at times, can be interpreted as binary digits as well. Note that when A is 0, Y is 1 while Y changes to 1 when A is made to change to 0. In other words, A and Y are mutually

complementary, which corresponds to the Boolean operation of *complementation*. One expresses this by writing Y as $Y = \overline{A}$.

TRUTH TABLES							
NOT		OR			AND		
A	Y	A	B	Y	A	B	Y
0	1	0	0	0	0	0	0
1	0	0	1	1	0	1	0
		1	0	1	1	0	0
		1	1	1	1	1	1

For the OR gate, A and B denote the two input variables while Y stands once again for the output variable. The value of Y for specific combinations of values of A and B are to be read off from the truth table, all the four possible combinations being shown. For instance, if any one of the input variables is 1 and the other is 0, the value of the output is 1. One expresses the relation between Y and A , B in the form $Y = A + B$, corresponding to the operation of *disjunction* in Boolean logic. The relation between the inputs (A , B) and the output (Y) for the AND gate, which can be similarly read off from the truth table, is expressed in the form $Y = A \cdot B$, and corresponds to the Boolean operation of *conjunction*.

As can be seen from the truth tables, both the AND and OR operations are commutative ($A + B = B + A$, $A \cdot B = B \cdot A$ for arbitrary values (0 or 1) assigned to A and B), and the AND operation is distributive with respect to the OR operation ($A \cdot (B + C) = A \cdot B + A \cdot C$).

The actual circuitry for each of these logic gates involves electronic components like resistances, diodes, transistors, and DC power supplies. For instance, a two-input AND gate or a two-input OR gate can be constructed with two diodes as the non-linear, or non-ohmic elements (i.e., components for which voltages and currents are not linearly related, such components being essential in electronic circuits in general, and in digital circuits in particular). A NOT gate, on the other hand, can be constructed with a n-p-n

transistor. However, the most commonly used components in the mass scale production of digital circuits are a particular variety of transistors termed MOSFETs.

19.7.2.1 The OR and AND gates with diodes

Fig. 19-25 illustrates the construction of (A) a two-input OR gate, and (B) a two-input AND gate with diodes.

In (A), the two inputs (in the form of *high* and *low* voltages, corresponding to Boolean 1 and Boolean 0 respectively) are imparted to the p-terminals of two p-n junction diodes D_1 , D_2 . The high voltage for either input may be taken from a DC voltage source supplying a fixed voltage (say, 5V), while the low input to either terminal may be realized by connecting that terminal to the ground (0 V). The two inputs, considered as Boolean variables, are denoted by A and B respectively. The output (represented by the Boolean variable Y) is taken from the common n-terminal of the two diodes, and appears as a high or low voltage (depending on whether $Y = 1$ or $Y = 0$) across an appropriate resistance R .

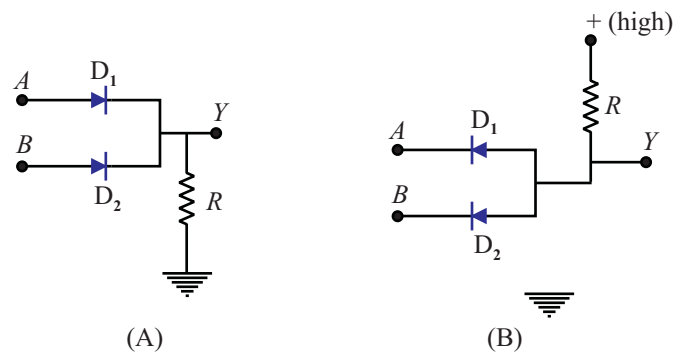


Figure 19-25: The OR and AND gates with diodes; (A) the OR gate; (B) the AND gate.

If both the inputs to the diodes are low (Boolean 0), i.e., both the p-terminals are grounded, then neither diode gets a forward bias, i.e., both diodes are non-conducting, and the output terminal, connected to the ground through R , attains the ground voltage, i.e., $Y = 0$. If, on the other hand any one of the diodes is given a high input then that diode becomes conducting, and the output terminal attains the same voltage as the

input terminal of that diode (with a relatively small voltage drop across the diode, which is $\approx 0.6\text{V}$ for a silicon diode.) In other words, the output Y is then 1. The same remains true if both the diodes receive a high input. In other words, $Y = 1$ if at least one of the two inputs A and B is 1. This, precisely, corresponds to the truth table of the two-input OR gate given in sec. 19.7.2.

In fig. 19-25(B), on the other hand, the input voltages are given to the n-terminals of the diodes D_1 and D_2 while the p-terminals are connected to each other to form a single common terminal, which is connected to the voltage source providing the fixed high voltage (say, 5V as in (A) above) through the resistance R (chosen of an appropriate value). The output voltage, corresponding to the Boolean variable Y (say) now appears between the common p-terminal and the ground (0V, corresponding to the low voltage representing Boolean 0).

In this set-up, if both the inputs are high, corresponding to Boolean 1, then the voltages across both the diodes are zero since the common p-terminal is also connected to the voltage source providing the same high voltage (5V in the present instance) and so, none of the diodes can be in the conducting state since the requisite cut-in voltage (see sec. 19.3.4) is not developed across either diode. In other words, the output terminal shows a high voltage (5V) with respect to ground, corresponding to $Y = 1$. If, on the other hand, at least one of the diodes receives a low input, then that diode becomes conducting and the voltage then appearing at the output terminal is then the same as this low input voltage (differing only by the relatively small diode drop $\approx 0.6\text{V}$), corresponding to $Y = 0$. These input-output relations reflect the truth table of the two-input AND gate given in sec. 19.7.2.

19.7.2.2 The NOT gate with a transistor

Fig. 19-26 depicts the circuit for a NOT gate with a n-p-n transistor (T). The single input (represented by the Boolean variable A) is fed to the base of the transistor through a base resistance R_B where the input terminal may be connected either to the ground (low voltage, $A = 0$) or to the positive terminal of a DC voltage source supplying a fixed

voltage (high, say, 5V, $A = 1$) with respect to the ground. The collector of T is connected to the same voltage source (5V in the present instance) through the resistance R_C . The output (Boolean variable Y) corresponds to the voltage level (high or low, represented by $Y = 1$ or $Y = 0$ respectively) between the collector and the ground.

The resistances R_B and R_C are so chosen as to cause the transistor to operate in the saturation mode when a positive voltage is applied to its base making the base-emitter diode forward biased.

Thus, with a high input ($A = 1$), one will have $V_{CE} \approx 0$ (more precisely, V_{CE} will lie in the range from 0 to 0.2V), i.e., $Y = 0$ (low). On the other hand, with a low input $A = 0$, the base-emitter diode will fail to be forward biased and the transistor will remain in cut-off. This implies that $I_C = 0$, and that there will be no potential drop across R_C . In other words, the high voltage level will appear at the output ($Y = 1$). The circuit thereby realizes the truth table of the NOT gate (refer to sec. 19.7.2).

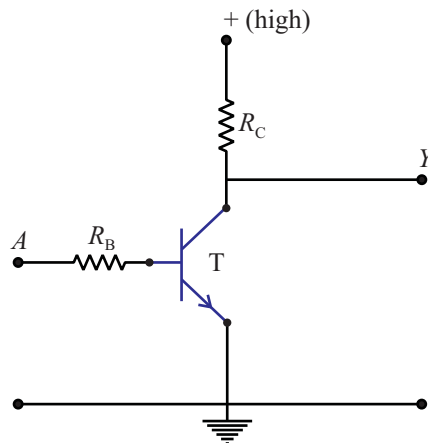


Figure 19-26: The NOT gate with a transistor; by a proper choice of the resistances R_B and R_C , the transistor is made to switch between the saturation and the cut-off modes of operation as the input voltage is made *high* or *low* respectively.

19.7.3 Logic families

Logic gates and logic circuits of various descriptions are available these days in the form of *chips*, which are mass-produced units fabricated with the *integrated circuit* technology. In the case of the more simple logic circuits like, for instance, the elementary gates (NOT, OR, AND) or, say, the NOR and NAND gates (see sec. 19.7.4 below) a chip may contain several gates of a particular type while, for larger and more complex logic circuits, a chip may contain one single circuit.

The chips available in the market belong to a number of *families*, where a single family includes a large variety of chips, each variety corresponding to a logic circuit of a given description. Any given family is based on a certain technology of production, and all chips of that family share a number of common characteristics, being *compatible* with one another. However, chips of various different families are not necessarily compatible, though they share certain common industry standards.

Logic families can be grouped into those based on bipolar junction transistors (BJT) and the ones based on metal-oxide semiconductor (MOS) transistors. Diode-based families are almost obsolete in present day digital circuit technology. A few of the more widely accepted logic families are the TTL (transistor-transistor logic) family, the NMOS (n-channel MOS) family, and the CMOS (complementary MOS, making use of both n-channel and p-channel devices) family.

19.7.4 The Exclusive-OR, NOR, and NAND gates

Fig. 19-27(A, B, C) shows three logic gates termed the EXOR, NOR and NAND, while their truth tables are also shown below. These are non-elementary gates in the sense that their underlying logical operations are compositions of the basic operations represented by the NOT, OR, and AND gates, where a composition of more than one operations involves a successive application of these operations.

While the NOR and NAND gates can be made up from the elementary gates, they are *universal* gates in the sense that any of the elementary gates or, for that matter, the

circuit realization of *any* Boolean function, can be made up with NOR or NAND gates alone. The exclusive-OR (ex-OR, EXOR, or XOR in brief) also finds wide application in digital circuits.

Notice from the truth tables that the NOR gate performs an OR operation followed by a NOT operation. The output (Y) for this gate can then be expressed in terms of the inputs(A, B) as $Y = \overline{A + B}$. Similarly, the NAND gate performs an AND operation followed by a NOT operation, and one can write $Y = \overline{A \cdot B}$. The EXOR operation is the same as the OR operation with the sole difference that the output for $A = 1, B = 1$ is 0 instead of 1 : the EXOR produces the output 1 only if just one of the inputs has the value 1. One writes, for the EXOR gate, $Y = A \oplus B$.

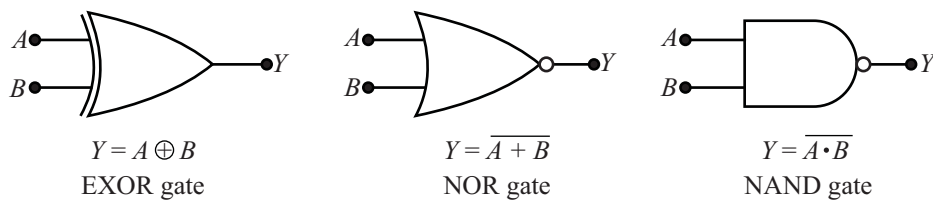


Figure 19-27: Circuit symbols for the (A) EXOR, (B) NOR, and (C) NAND gates, each with two inputs (A, B), and an output (Y); the Boolean equation expressing the output in terms of the input is shown in each case.

TRUTH TABLES								
EXOR			NOR			NAND		
A	B	Y	A	B	Y	A	B	Y
0	0	0	0	0	1	0	0	1
0	1	1	0	1	0	0	1	1
1	0	1	1	0	0	1	0	1
1	1	0	1	1	0	1	1	0

EXOR, NOR, and NAND gates with more than two inputs are also used. In each of these three cases the operation is *commutative* and *associative* (as is the case with OR and AND gates) in the sense that an interchange of the inputs does not change the output.

19.7.5 Boolean identities and Boolean expressions

19.7.5.1 De Morgan's identities

A *Boolean identity* involves two Boolean expressions, each depending on the same set of variables (say, A , B , C , ..., termed the input variables), such that both the expressions assume identical values (0 or 1) for any and every choice of values of the input variables.

Two identities of fundamental importance in digital electronics are the *De Morgan identities*,

$$\overline{A + B} = \overline{A} \cdot \overline{B}, \quad (19-18a)$$

and

$$\overline{A \cdot B} = \overline{A} + \overline{B}. \quad (19-18b)$$

In eq. (19-18a), for instance, assigning the values 0 and 1 to the variables A and B respectively, one finds that each of the expressions $\overline{A + B}$ and $\overline{A} \cdot \overline{B}$ assumes the value 0 (check this out), and a similar equality holds for each of the other three possible combinations of values of the input variables. In the same manner, eq. (19-18b) also represents a valid identity.

Consider, on the other hand, the expressions $A \oplus B$ and $A + B$. These two expressions assume different values for the input combination $A = 1$, $B = 1$, though their values agree for the other three possible combinations of values of the input variables. Hence a proposed relation of the form $A \oplus B = A + B$ (?) will not count as a valid Boolean identity.

While (19-18a), (19-18b) are powerful identities, a few other useful Boolean identities are as follows. In these identities, A can be any of the two Boolean constants 0 and 1. In other words, it represents a Boolean variable.

$$1 + A = 1, \quad 1 \cdot A = A, \quad (19-19a)$$

$$0 + A = A, 0 \cdot A = 0, \quad (19-19b)$$

$$A + A = A, A \cdot A = A, \quad (19-19c)$$

$$A + \overline{A} = 1, A \cdot \overline{A} = 0. \quad (19-19d)$$

These can be made use of in simplifying Boolean expressions, and in establishing other Boolean identities.

Problem 19-7

Establish the Boolean identity $(A + B) \cdot (\overline{A} + C) = A \cdot C + \overline{A} \cdot B$, where A, B, C are any three Boolean variables.

Answer to Problem 19-7

HINT: $(A + B) \cdot (\overline{A} + C) = AC + \overline{A}B + BC = AC + \overline{A}B + (A + \overline{A})BC = AC(1 + B) + \overline{A}B(1 + C) = AC + \overline{A}B$.

Here the basic Boolean identities referred to above have been made use of. The AND operation is commonly expressed in an abbreviated form by writing, for instance, AB in place of $A \cdot B$.

19.7.6 The binary numbers. Binary arithmetic

19.7.6.1 The Decimal, Binary, Octal, and Hex systems

The numbers we commonly use are *decimal* numbers, and are made up of the digits 0, 1, 2, ..., 9. Digital electronics, on the other hand, makes use of *binary* numbers made up of just two digits, namely, 0 and 1. While the *base* in the decimal system is 10, that in the binary system is 2. For instance,

$$(1011.01)_2 = 1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 + 0 \times 2^{-1} + 1 \times 2^{-2} = (11.25)_{10},$$

where the base has been explicitly denoted with the help of a suffix.

Considering the representation of a number in terms of a given base, each digit in the number is to be multiplied with the base raised to some power or exponent. For digits to the left of the decimal point, the exponent increases by 1 for the successive digits, starting from 0 for the first digit to the left. Similarly, for the digits to the right of the decimal point, the exponent decreases by 1 for the successive digits starting from -1 for the first digit to the right. The values so obtained for all the digits are to be added to arrive at the value of the number as represented with the help of the given base.

This rule applies for representations other than the binary and the decimal ones, among which the octal and the hex representations deserve mention. The *octal* system, for instance, uses the base eight, with digits 0, 1, ..., 7 as the basic digits, while the *hexadecimal* (*hex* in short) system uses the base sixteen, with 0, 1, 2, ..., 9, A, B, C, D, E, F as the basic digits. Thus, for instance,

$$(C)_{16} = (12)_{10} = (1100)_2, (21C)_{16} = (540)_{10} = (1034)_8,$$

(check these two statements out).

Hex numbers are often identified with the symbol H attached at the end instead of the suffix 16. Thus, for instance,

$$9FH = (9F)_{16} = (159)_{10}.$$

While in common usage we mostly employ decimal numbers without explicit mention of the base 10, in digital electronics one mostly uses binary numbers which are strings made up of 0's and 1's, without explicitly displaying the base 2.

19.7.6.2 Bits and Bytes

A single digit in a binary number is termed a *bit*. Binary numbers are often stored in units of 8 bits taken at a time, termed *bytes*. Thus, for instance, the binary number 10110 contains five significant bits, while a byte in which the same number is stored would be 00010110. Here the three 0's to the left are not significant digits, being brought

in to make up the 8 bits in the byte.

However, the way the binary numbers are stored in digital systems, the left-most digit in the above byte would be significant in a broader sense: it would tell us that the number under consideration is a non-negative one (see sec. 19.7.7.1).

19.7.7 Eight-bit arithmetic

19.7.7.1 1's complement and 2's complement

The complement (1's complement or 2's complement) of a binary number is used in digital arithmetic to represent numbers with a *negative sign* and to implement the process of *subtraction*.

The 1's *complement* of a binary number is obtained by replacing 1's with 0's and 0's with 1's. For instance, the 1's complement of the 8-bit number

$$10001101$$

is

$$01110010.$$

Note that the left-most zero in the second number is not to be omitted since in computer storage every bit in a byte must be filled up with either a 0 or a 1.

However, 1's complement is rarely used in actual computer storage where another representation, namely the 2's *complement* is employed universally. The 2's complement of a binary number is obtained by first taking its 1's complement and then adding 1 to it. Thus, for instance, the 2's complement of the 8-bit binary number 10001101 is 01110011.

Whenever a *negative* number is to be stored in a computer, it is stored in the 2's complement form. We illustrate this in the framework of *8-bit arithmetic*, which was used by early computers in storing numbers and performing arithmetical operations on those.

In 8-bit arithmetic a byte is used in storing integers ranging from $(-128)_{10}$ to $(+127)_{10}$. All the non-negative integers within this range are stored in their direct binary forms, with all the eight bits filled up (with 0's to the left, if necessary). For instance, the integer $(95)_{10}$ is stored as 01011111. In this representation of non-negative integers, a 0 always appears in the left-most bit, termed the *most significant bit* or, in short, the *MSB*. The bit in the extreme right position is similarly termed the *least significant bit*, or *LSB*. The largest positive integer to be stored in a byte is $(127)_{10}$, i.e., 01111111.

The *negative* integers, on the other hand, are stored by converting their magnitudes into 2's complements. Thus, for instance, $(-71)_{10}$ is stored as 10111001 because this 8-bit string represents the 2's complement of $(71)_{10}$, the magnitude of the negative integer in question.

This, then, is the format for storing integers ranging from $(-128)_{10}$ to $(+127)_{10}$ in 8-bit arithmetic.

In this format, the MSB for a negative integer always happens to be 1, the negative number with the largest magnitude, namely $(-128)_{10}$ being represented by 10000000.

19.7.7.2 Addition and subtraction in 8-bit arithmetic

Keeping in mind this format of storing numbers in 8-bit arithmetic, one can now address the question as to how *addition* and *subtraction*, the two basic arithmetical operations, are performed by a computing system.

To start with, the two numbers on which the operation is to be performed (each of the two arithmetical operations is always performed on two numbers taken at a time), are stored in the above format. In the case of addition, these two are added up with a piece of digital circuitry referred to as the *adder* whose output gives the sum of the two numbers, once again in the above format.

In the case of *subtraction*, on the other hand, the subtrahend (i.e., the number to be subtracted) is first converted into its 2's complement form. The two resulting numbers

(the minuend and the 2's complement of the subtrahend) are then *added up* with the help of the adder. The output of the adder then gives the required result, once again in the above format.

As an example, consider the subtraction $(33)_{10} - (-65)_{10}$. Here the minuend, being a positive number, is stored in its binary form as

00100001.

The subtrahend, on the other hand, being a negative number, is stored in the 2's complement form as

10111111.

Since this last string constitutes the subtrahend, it is to be converted into its 2's complement according to the rule of the game, to give

01000001

(which, in fact, is the string obtained by the direct binary conversion of $(+65)_{10}$).

The adder circuit is now called upon to perform addition of the two strings 00100001 and 01000001. On receiving its inputs, it yields as output the string

01100010,

which is the result of the subtraction problem stored in the above format. In the present instance, the result is $(+98)_{10}$ and its stored form would be simply the string obtained by direct binary conversion, which is precisely the string 01100010.

19.7.7.3 Overflow and carry

Two special situations that can arise in the above scheme of operations are (i) an *overflow*, and (ii) a *carry* in the MSB.

Overflow in 8-bit arithmetic is said to occur when the result of an operation lies outside the range of numbers $-(128)_{10}$ to $+(127)_{10}$ admissible in the 8-bit representation. This shows up as an incorrect sign (the binary digit in the MSB) for the number under consideration. For instance, if one attempts to work out the subtraction problem $(119)_{10} - (-69)_{10}$ in the above scheme of 8-bit arithmetic, one ends up with the incorrect answer 10111100, since the sign bit (1 in the MSB) indicates that the result is a negative number which, in fact, it is not. This is an example of an overflow where the actual result of the desired operation $((188)_{10})$ is outside the range of acceptability of 8-bit arithmetic.

In some situations a *carry* may appear at the MSB of the result of an operation. If this occurs without an overflow, the computing system simply ignores this carry, while still arriving at the correct result. If, however, the carry at the MSB occurs *simultaneously* with an overflow then simply ignoring the carry will not do because an overflow indicates a real problem, namely, that the capacity of 8-bit arithmetic has been reached.

Even early computing systems based on 8-bit arithmetic had to handle numbers far exceeding the range admissible in 8-bit arithmetic. For this, they had to use additional bytes for storing numbers and operating upon those. All operations were, however, performed byte-wise in accordance with the above rules of 8-bit arithmetic. The problem of overflow at the 8th bit of each byte had to be dealt with appropriate instructions applied to the system.

19.7.7.4 Binary multiplication and division

In binary multiplication, the basic operation is that of the AND gate since the truth table of the AND gate corresponds to the multiplication table of two binary digits ($0 \times 0 = 0$, $0 \times 1 = 1 \times 0 = 0$, $1 \times 1 = 1$). In a multiplication of two numbers with more than

one digits, the process of bit-wise multiplication (reproduced by the AND operation) is employed along with the algorithm of ‘repeated left-shift and add’ method used in the familiar multiplication algorithm of decimal (i.e., base 10) numbers.

In binary division, on the other hand, a computing system makes use of the algorithm of ‘repeated right shift and subtract’ process commonly employed in the division of decimal numbers. At every step of the process the digit appearing in the quotient (a 0 or a 1) is determined by *comparing* the subtrahend and the minuend.

19.7.8 The adder

The adder is a logic circuit that performs the addition of three binary digits, producing *two* outputs - a *sum* and a *carry-out*. It is the basic unit employed in the arithmetical processes performed by a digital computing system.

In the addition of two binary numbers, the computing system performs a bit by bit addition where, at each bit position, one has to actually add *three* binary digits, of which two come from the numbers to be added while the third appears as the *carry-in* (0 or 1) from the previous bit place. Denoting the Boolean variables corresponding to these three binary digits by A , B , and C respectively, and those corresponding to the binary digits for the sum and the carry-out by Y and C' (say), the truth table expressing Y and C' in terms of A , B , and C is given below.

The truth table can be explained by referring to any one row in it, say, the fourth one from the top. In terms of binary digits, one has $0 + 1 + 1 = 10$, where the sum, i.e., the digit appearing in the bit place under consideration is seen to be 0, while the carry-out is 1. Hence, the relevant values of the Boolean outputs are $Y = 0$, $C' = 1$.

ADDER				
A	B	C	Y	C'
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1

The entire truth table for the adder is represented in a compact form by the following functional relations between the output and the input variables.

$$Y = A \oplus B \oplus C, \quad C' = A \cdot B + B \cdot C + C \cdot A, \quad (19-20)$$

where \oplus denotes the EXOR operation.

Based on these relations, the logic circuit for the adder can be constructed.

For the sake of completeness, I may mention the *half-adder*, which is a logic circuit designed for the addition of *two* binary digits, which produces, with any given pair of inputs, a sum and a carry-out as in the adder. The half-adder is relevant, in an addition job involving two binary numbers, in the summing up of the LSB's of the two numbers, since this summing up does not involve a carry-in. However, this task can be performed equally well with an adder for which the carry-in has been set at 0.

In a digital computing system the addition and subtraction jobs (including those involved in multiplication and division) are performed by *adders in succession*, there being eight such adders necessary in 8-bit arithmetic. However, I will not enter here into a more detailed consideration of the logic circuits necessary for such arithmetical computations.

19.7.9 Flip-flops

The adder and the half-adder are instances of combinational logic circuits where the values of the output variables are determined by the *current* values of the input variables, where the dependence between the two sets of variables can be expressed in terms of Boolean functions. Any such Boolean function can be realized with elementary logic gates (or with universal ones) in more ways than one. A few combinational circuits commonly used in digital systems are the arithmetic circuits such as the half-adder and the adder, the *multiplexer* and *demultiplexer*, and the *encoder* and *decoder*.

In contrast to the operation of a combinational circuit, the values of the output variables of a *sequential* circuit are determined not only by the current values of the input variables but by a sequence of values depending on *past* states of the circuit as well. In other words, a sequential circuit operates with a *memory*. Sequential circuits constitute the backbone of large scale digital circuits of enormous functional capacity, and are made up of elementary components termed *flip-flops*. Flip-flops, in turn, are built with the elementary (or, more commonly, with the universal) logic gates by the appropriate use of *feedback* connections between these gates.

19.7.9.1 the *SR* flip-flop

As a simple instance, fig. 19-28 depicts a *SR flip-flop* made with a pair of NOR gates where the output of each NOR gate is fed back to one of the inputs of the other gate. The two remaining input terminals of the set-up are termed the *set* and the *reset* terminals, receiving input data represented by variable S, R , each of which can have values 0 (*low*) or 1 (*high*). The Boolean variables corresponding to data at the two output terminals of the set-up are complementary to each other and are denoted by Q, \bar{Q} (such complementary outputs are common in most flip-flop set-ups).

A close analysis of the circuit tells us that the input and the output variables satisfy,

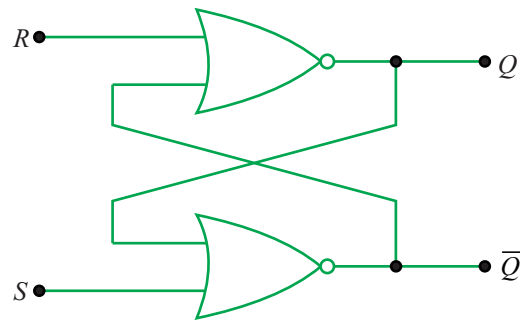


Figure 19-28: A *SR* flip-flop made up of two NOR gates with feedback; data represented by Boolean variables S and R are fed in the form of pulses to the two input terminals termed *set* and *reset*; when a pulse is fed to the set terminal, the output Q attains the value 1 regardless of its previous value while, likewise, a pulse fed to the reset terminal results in $Q = 0$; when no data are fed to either terminal, Q continues in its previous value: the flip-flop retains its memory; simultaneous feeding of pulses to both the input terminals ($S = 1, R = 1$) results in an indeterminate output.

between them, the *implicit* relation

$$Q = S\bar{R} + \bar{R}Q, \quad (19-21)$$

for which the truth table is shown in below.

SR flip-flop		
S	R	Q
0	0	<i>last value</i>
1	0	1
0	1	0
1	1	?

One observes that, with $S = 0, R = 0$, Q can be either 0 or 1, which means that, when the set and reset are both at *low* level, the output Q continues at its previous value: the output of the flip-flop does not change. When the data at the input terminals are $S = 1, R = 0$ (this is achieved by sending an input *pulse* to the set terminal, see below), the output gets *set* to $Q = 1$, while with an input pulse fed to the reset terminal $S = 0, R = 1$, the output is *reset* to $Q = 0$. As seen from the truth table, when neither of the input terminals receives a data pulse ($S = 0, R = 0$), the output continues at its last

value ($Q = 1$ if the last pulse received was at the set terminal, and $Q = 0$ if the last pulse received was at the reset terminal): the set-up retains its memory unless it is set (when Q becomes 1 regardless of its last value) or reset (Q becomes 0 regardless of its last value).

What happens when *both* the input terminals are *high* ($S = 1, R = 1$)? This is not a standard input configuration for the SR flip-flop and is avoided in practice, because then the two outputs are not complementary to each other, *both* attaining the value 0. The state of the flip-flop after the passage of the pulses (when both the input terminals attain a *low* level) becomes *indeterminate*: if the passage of the R -pulse precedes that of the S -pulse even by an extremely short interval of time, then the output will get set at $Q = 1$ while, if the S -pulse subsides first, the output will remain at $Q = 0$.

As mentioned above, data are fed to the input terminals in the form of *pulses* of short duration, with the voltage level becoming *high* (Boolean 1) for a short duration and then subsiding to *low* (Boolean 0), as in fig. 19-29(A). Pulses arrive at any given terminal in succession, and the intervening time intervals correspond to the *low* voltage level. In fig. 19-29(B), a sequence of pulses is shown, where the entire profile is assumed to propagate from left to right with the passage of time. each pulse consists of a *leading edge* and a *trailing edge*, corresponding to the earlier and the later of the two transitions of the voltage level.

Though it is unlikely that, during the operation of the flip-flop, two pulses will arrive at the input terminals simultaneously, still such a coincidence cannot be ruled out, which is why, the technique of *edge triggering* is commonly employed in flip-flop operation. A pulse is made to pass through a *differentiating* circuit wherein its two edges give rise to two extremely sharp pulses, one of which is made effective in setting or resetting the flip-flop.

You will find the circuit symbol of a SR flip-flop in fig. 19-30 where the set-up for a D flip-flop is shown.

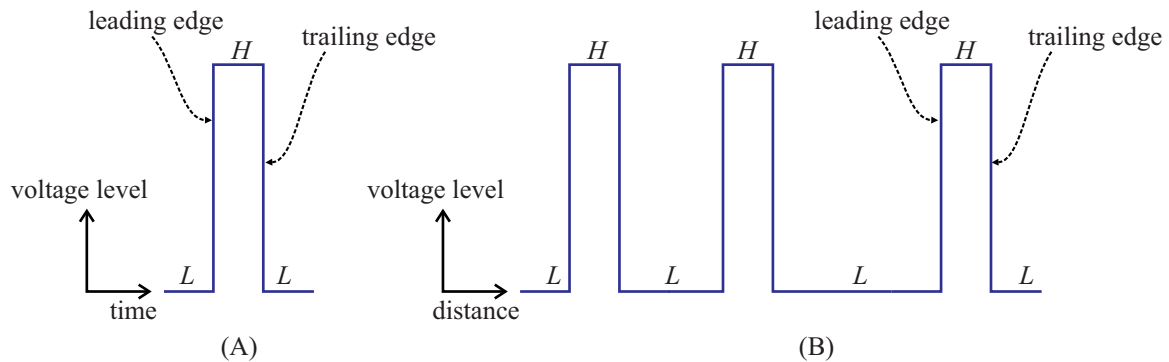


Figure 19-29: (A) Depicting a rectangular pulse with *low* (L) and *high* (H) voltage levels, and with the *leading* and *trailing edges*; the horizontal axis corresponds to increasing time, while the location of the pulse is imagined to remain unchanged; (B) a sequence of pulses, with successive pulses separated from each other, forming a profile at some specified instant of time; the profile is assumed to propagate from the left to the right with the passage of time (dotted arrow); leading and trailing edges are shown.

19.7.9.2 the *D* flip-flop

While the *SR* flip-flop is triggered by data fed to two terminals, the *D* (or *delay*) flip-flop stores data fed to only one terminal, the storing being *enabled* or *disabled* by means of *clock pulses* fed to a second terminal. Fig. 19-30 depicts the set-up of a *D* flip-flop involving two AND gates and a *SR* flip-flop (represented by circuit symbol on the right of the figure). In the figure, *D* stands for the Boolean variable representing the input data - two of the input terminals of the set-up receive data *D* and \overline{D} respectively, the latter as the output of a NOT gate. As for the remaining two input terminals, both receive the clock pulse represented by *CLK* that arrives in succession, each pulse raising the voltage level to *high* ($\equiv 1$) for a short duration and then causing it to subside to *low* ($\equiv 0$), thereby enabling and disabling the operation of the flip-flop.

During the short time interval in which the circuit is enabled by $CLK = 1$, any data *D* fed to the input terminal is transferred to the output so as to make $Q = D$. As the clock pulse subsides ($CLK = 0$), the circuit gets disabled, and data fed at the input terminal does not have any effect on *Q*: the circuit ‘remembers’ the last value attained during the activation by the last clock pulse, to be activated once again by the next clock pulse. The truth table of the *D* flip-flop is given below, showing the enabling and disabling action

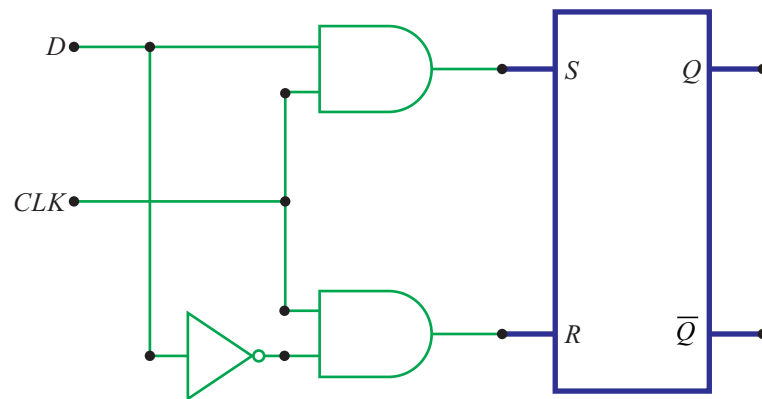


Figure 19-30: Depicting the set-up of a D flip-flop with two AND gates and a SR flip-flop, the box on the right being the circuit symbol of the latter; one input each of the two AND gates receives the clock pulse (Boolean variable CLK) that enables and disables the transfer of data D to the output Q ; the remaining terminals of the AND gates receive data D and \bar{D} , the latter through a NOT gate; as CLK becomes 1, the data D is transferred to the output; for $CLK = 0$, the output Q remains at its last value.

of CLK .

D flip-flop		
CLK	D	Q
0	<i>anything</i>	<i>last value</i>
1	0	0
1	1	1

With clock pulses of short but finite duration (*level-triggering*), more than one data bits may arrive at the input terminal during this short interval, causing unwanted data transfer to the output. This is remedied in the *edge-triggered* D flip-flop by differentiating the clock pulses and sending in the sharp spikes of extremely short duration resulting from the differentiation of one of the two edges of each pulse. Fig. 19-31 depicts the logic symbol of (A) a *positive* edge-triggered and (B) a *negative* edge triggered flip-flop. In the former, the *upward spike* resulting from the leading edge of the clock pulse by the process of differentiation is made to enable the flip-flop, while in the latter it is the downward spike that does the job.

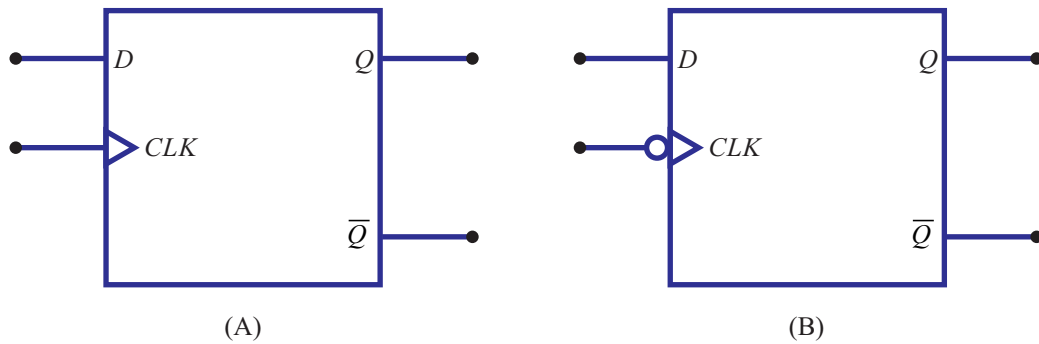


Figure 19-31: Logic symbol of (A) a positive edge-triggered and (B) a negative edge-triggered D flip-flop; in (B), note the bubble for the clock input, which is a symbol for negation.

19.7.9.3 The JK flip-flop

The JK flip-flop is widely used in digital systems in *counters*. Fig. 19-32 depicts a realization of the JK flip-flop with the help of two 3-input AND gates and a SR flip-flop. The clock pulse enabling the unit is made to pass through a differentiating circuit (not shown; the differentiating circuit is made of a resistance-capacitance combination of very low time constant) for the realization of edge-triggering.

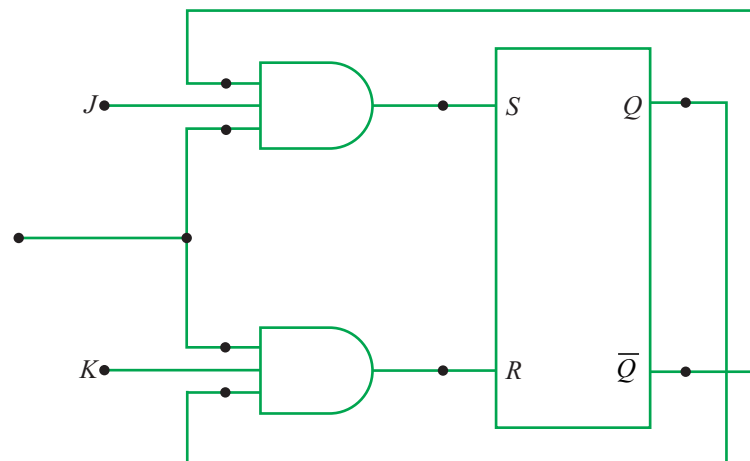


Figure 19-32: Depicting the set-up of a JK flip-flop with two 3-input AND gates and a SR flip-flop, the box on the right being the circuit symbol of the latter; one input each of the two AND gates receives the clock pulse (Boolean variable CLK) that enables and disables the transfer of data D to the output Q ; the remaining input terminals of the AND gates receive data J, K and, additionally, are connected to the output terminals of the system (Q, \bar{Q}) in feedback loops; the outputs of the AND gates provides the S and R inputs of the SR flip-flop.

As shown in the figure, the clock pulse is fed to one input of each of the two AND gates, and the input data (referred to as J, K) are also presented to one input each of the two, while the remaining inputs of the AND gates are connected to the output terminals of the SR flip-flop in feedback loops. The functioning of the entire set-up can be described as follows.

First, *in the absence of a clock pulse, the flip-flop is disabled*: it just latches on to its last configuration whatever be the values of J and K .

On the arrival of a clock pulse, on the other hand, the system makes a response that is *controlled* by the values of J, K at the time of the arrival of the clock pulse.

To begin with, for $J = 0, K = 0$, nothing interesting happens: the system makes *no response* to the clock pulse, and the output (Q, \overline{Q}) continues to remain at its previous value.

For $J = 1, K = 0$, the flip-flop gets *set* on the arrival of the clock pulse, giving $Q = 1, \overline{Q} = 0$.

Likewise, with $J = 0, K = 1$, the the arrival of a clock pulse *resets* the flop-flop, resulting in $Q = 0, \overline{Q} = 1$.

Finally, with $J = 1, K = 1$ (recall that such a combination of inputs is avoided in a SR flip-flop), the system makes a response to the clock pulse by *toggling over*: if the last output was $Q = 1, \overline{Q} = 0$, it now becomes $Q = 0, \overline{Q} = 1$, and *vice versa*. Thus, the JK flip-flop is, in a sense, a combination of the SR flip-flop and the T flip-flop where the latter is an elementary memory unit like the D flip-flop, with the difference that, when triggered by a clock pulse, the output either remains the same (for input $T = 0$) or toggles over (for $T = 1$).

At times, it becomes necessary to set or reset a flip-flop independently of the input data or the clock pulse. This is achieved by making use of two additional data terminals referred to as the *preset* and the *clear*.

The truth table of a JK flip-flop is given below. In this truth table Q denotes the output before the flip-flop makes its response, while Q' is the output resulting from the response. The upward arrow represents the positive edge of the clock pulse (assuming positive edge-triggering), while an X in the column under CLK means that it is a 'don't care' input: the system makes no response to the clock pulse.

JK flip-flop			
CLK	J	K	Q'
X	0	0	Q
\uparrow	0	1	0
\uparrow	1	0	1
\uparrow	1	1	\overline{Q}

Fig. 19-33 depicts the logic symbol of a positive edge-triggered JK flip-flop with preset (PR) and clear (CLR).

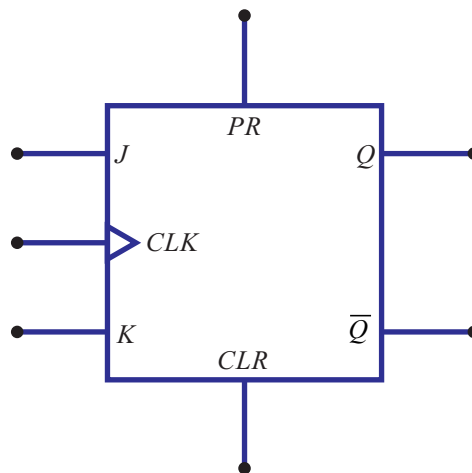


Figure 19-33: Logic symbol of a positive edge triggered JK flip-flop with preset and clear.

The simple JK flip-flop described above suffers from a problem referred to as 'racing' which essentially is a consequence of the finite width, however small, of the clock pulse or spike (in the case of edge triggering), as a result of which the unit works as desired

only with very high frequency clocks. Such high frequency is a disadvantage in many digital systems, as a result of which the 'master-slave' JK (or the $MSJK$ in brief) is almost universally used as the basic flip-flop unit. Such a flip-flop is made up of two JK units (master-slave set-ups with two SR units are also common), where the output terminals of the first of the two (the 'master', to which the J and K inputs to the system are fed) are connected to the input terminals of the second (the 'slave'), the output terminals of the latter being connected in feedback loops to the inputs of the 'master'. Most importantly, the 'master' and the 'slave' are triggered by opposite edges of the *same* clock pulse, as a result of which there occurs a delay between the outputs of the two, with the 'slave' copying the 'master'.

Applications of flip-flops, the basic units of sequential logic circuits, cover a vast area. Among these, *registers* (also referred to as *shift registers*) constitute a major field of application. A register is designed to store a number of data bits, in which bits can be moved in and moved out either serially or in parallel.

Another area of application, of enormous importance in the field of digital systems, is that of *counters*. A counter is a device that keeps a tally of clock pulses sent to it. Built with flip-flops as the basic units, it operates in a manner similar to a register, differing from the latter in one significant respect: a counter has a characteristic internal sequence of states through which it passes as successive clock pulses are fed to it, while a register has no such characteristic sequence. The states of a register are determined by the data fed to it - but a counter receives no external data other than the series of clock pulses it receives, and possibly some *control* data.

A counter can belong to either of two broad categories - *asynchronous* and *synchronous*. An asynchronous counter (also known as a 'ripple counter') is one in which the flip-flops do not receive their triggering pulses simultaneously - instead, they are triggered sequentially, with the output of one flip-flop providing the clock pulse for the next, introducing a small *delay* between the triggering events of successive flip-flops.

In a synchronous counter, on the other hand, all the flip-flops making up the counter

receive their triggering pulses *simultaneously* since the *same* clock pulse triggers all of those.

Bibliography

- [1] Lord Rayleigh, *The Principle of Similitude*, Nature, vol 95, p 66-68 (1915).
- [2] E. Buckingham, *On Physically Similar Systems; Illustrations of the Use of Dimensional Analysis*, Physical Review, vol. 4, p 345-376 (1914).
- [3] G. W. Bluman and S. Kumei, *Symmetries and Differential Equations*, Springer-Verlag, New York (1989).
- [4] D. F. Young, B. R. Munson, T. H. Okiishi, and W. W. Huebsch, *A Brief Introduction to Fluid Mechanics*, Fifth Ed., John Wiley & Sons, Inc. (2011).
- [5] R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman lectures on Physics*, vol.2, Narosa Publishing House, New Delhi (originally published 1964; reprint, 1995).
- [6] A. A. Sonin, *The Physical basis of Dimensional Analysis*, Department of Mechanical Engineering, MIT, Cambridge (2001) [available at URL: web.mit.edu/2.25/www/pdf/DA_unified.pdf (visited 17 August, 2017)].
- [7] Bharat Bhushan, *Introduction to Tribology*, John Wiley & Sons, Ltd., West Sussex, U.K., 2nd edition (2013).
- [8] K.C. Ludema, *Friction, Wear, Lubrication: A text-book in tribology*, CRC Press, Boca Raton (1996).
- [9] J. W. Tester and M. Modell, *Thermodynamics and Its Applications*, Prentice Hall PTR, New Jersey, 3rd edition (1997).

- [10] Peter Behroozi, Kimberly Cordray, William Griffin, and Feredoon Behroozi, *The calming effect of oil on water*, American Journal of Physics, vol 75, p 407-414 (2007).
- [11] I. Galili and E. Goihbarg, *Energy transfer in electrical circuits: a qualitative account*, American Journal of Physics, vol. 73, p 141-144 (2005).
- [12] S.L. O'Dell and R.K.P. Zia, *Classical and semi-classical diamagnetism: A critique of treatment in elementary texts*, American Journal of Physics, vol. 54, p 32-35 (1986).
- [13] M. Huemer, *Correct Resolution of the Twin paradox*, available at <http://home.earthlink.net/owl232/twinparadox.pdf> (visited 12 August, 2017).
- [14] J. Natario, *General Relativity Without Calculus*, Springer-Verlag, Berlin (2011).

Index

C-R circuit

- decay of charge in, 1301
- growth of charge in, 1299
- time constant of, 1301

D flip-flop, 1882

I-V characteristic

- junction diode, 1819
- Zener diode, 1827

L-C-R circuit

- oscillations in, 1302
- with AC source, 1313

L-R circuit

- decay of current in, 1295
- growth of current in, 1292
- time constant of, 1294

T flip-flop, 1885

JK flip-flop, 1884

SR flip-flop, 1879

1's complement, 1873

2's complement, 1873

aberration, 884, 967

- and diffraction, 972
- chromatic, 969
- monochromatic, 892, 968

absolute scale of temperature, 672

absorption coefficient, 751

AC

- complex representation of, 1310
- mathematical description of, 1308
- phase in, 1309
- resonance in, 1317

AC *L-C-R* circuit, 1313

AC circuit

- analysis of, 1321

acceleration

- of a particle, 95
- angular, 203
- centripetal, 204, 209
- radial and cross-radial, 211
- transformation of, 123

acceleration due to gravity, 395

- and earth's rotation, 403
- variation of, 396

acceptor, 1808

acoustic pressure, 767

acoustic wave

- energy density, 811
- nonlinear, 817
- strain in, 803
- velocity of, 804

acoustic waves

coherence of, 832
 action-at-a-distance, 194
 adder, 1877
 adiabatic enclosure, 591
 adiabatic process, 595, 605
 work in, 640
 adiabatic work
 significance of, 607
 air column
 standing wave in, 838
 alpha decay, 1748
 ampere
 definition of, 7
 Ampere's circuital law, 1201
 applications of, 1203
 amplifier
 voltage gain, 1850
 AND and OR gates
 with diodes, 1864
 angle of contact, 564
 Young formula for, 566
 angular acceleration, 203
 angular momentum, 214
 about a point, 214
 about an axis, 215
 conservation of, 223
 in circular motion, 216
 angular motion, 196, 222
 angular oscillations, 335
 angular velocity
 about a point, 196
 about an axis, 199
 anode and cathode, 1111
 anomalous expansion
 and marine life, 697
 of water, 696
 anti-particle, 1771
 aperture
 for spherical mirror, 922
 Archimedes' principle, 486
 area
 as vector, 55
 areal velocity, 411
 asperities
 role in friction, 297
 asynchronous motor, 1277
 atmospheric pressure, 484
 standard, 485
 atom
 elliptic orbits in, 1691
 single-electron states, 1689
 atomic energy shell, 1701
 atomic mass, 1687
 atomic mass unit, 20
 atomic nucleus, 1687
 atomic number, 1688, 1725
 atomic spectra, 1723
 atomic subshell, 1701
 atomic volume, 1687
 attenuation
 in conducting medium, 1397
 aufbau principle, 1702

Avogadro number, 625
 Avogadro's law, 638
 axial vector, 54, 219
 azimuthal quantum number, 1694
 back EMF
 in DC motor, 1281
 in inductor, 1288
 in transformer, 1336
 bar and torr, 485
 baryon number, 1770
 baryons, 1767
 base units, 4
 basic interactions
 conservation principles, 1774
 beats, 846
 bel, 815
 Bernoulli's principle, 511
 in the working of siphon, 520
 beta decay, 1750
 energy distribution in, 1752
 bimetallic strip, 690
 binary arithmetic, 1871
 binary numbers, 1861, 1871
 binding energy, 1582
 of nucleus, 1735
 surface correction, 1739
 binding energy curve, 1737
 binormal, 210
 bit, 1872
 BJT, 1833
 black body
 distribution function, 1561
 black body radiation, 746, 1560
 black holes, 1683
 blocking capacitor, 1849
 blue of the sky, 1523
 Bohr's postulates, 1569
 Bohr's theory, 1566
 and quantum theory, 1572
 and X-ray spectra, 1719
 applications of, 1576
 boiling point, 714
 Boltzmann formula
 for entropy, 663
 for probability, 662
 Boltzmann's constant
 unit of, 15
 Boolean algebra, 1860
 Boolean identities, 1869
 bosons and fermions, 1769
 bound systems
 and standing waves, 1577
 in quantum theory, 1551
 boundary friction, 304
 boundary layer, 538
 for flat plate, 540
 laminar, 539
 near curved surface, 543
 turbulent, 542
 boundary layer separation, 543, 546
 Boyle's law, 638
 Brewster angle, 1499

Brownian motion, 644
 Buckingham pi theorem, 26
 bulk modulus, 461
 adiabatic, 807
 bulk strain, 431
 bulk stress, 441
 bulk viscosity, 528
 buoyancy, 486
 center of, 489
 bypass capacitor, 1849
 byte, 1872
 calorimetry, 700
 fundamental principle of, 709
 camera, 975
 candela
 definition of, 8
 capacitance, 1070
 coefficients of, 1071
 of circular disk, 1087
 capacitor, 1070
 cylindrical, 1089
 energy of, 1092
 parallel plate, 1084
 spherical, 1078
 with dielectrics, 1098
 capacitors
 series and parallel, 1094
 capillary rise, 571
 Carnot cycle, 668
 T-S diagram for, 671
 Carnot engine, 669
 with an ideal gas, 669
 cell
 Daniell, 1108
 electrochemical, 1106
 electrolytic, 1110
 Galvanic, 1108
 primary and secondary, 1112
 center of buoyancy, 489
 center of curvature, 209
 center of gravity, 398
 center of mass, 159
 determination of, 163
 momentum of, 162
 motion of, 165
 of hemisphere, 164
 position of, 159
 velocity of, 162
 center of mass frame, 167
 in collision, 176
 centrifugal force, 127, 264
 centripetal acceleration, 204, 209
 change of state, 712
 latent heat in, 714
 charge
 unit of, 11
 charge density, 1002, 1053
 charged sphere
 energy of, 1047
 Charles' law, 638
 circuit symbol
 elementary gates, 1862

- EXOR gate, 1868
- junction diode, 1822
- light emitting diode, 1831
- transistor, 1834
- universal gates, 1868
- Zener diode, 1827
- circular motion, 201
 - angular momentum in, 216
 - uniform, 203
- circular polarization, 1381, 1495
- circular wire
 - magnetic field of, 1194
- co-planar forces
 - reduction of, 284
- coefficient of restitution, 175
- coefficient of thermal expansion, 698
- coefficient of viscosity
 - unit of, 11
- coefficients of capacitance, 1071
- coercive field, 1245
- coexistence curve, 719
- coherence, 832, 1405
- coherence and incoherence
 - in diffraction, 1480
- coherence length, 1442
- coherent source, 1500
- cold hardening, 451
- collisions
 - center of mass frame in, 176
 - elastic and inelastic, 170, 173
 - energy balance in, 172
 - head-on, 177
 - in planar motion, 181
 - momentum balance in, 174
- colour of thin films, 1456
- combinational logic, 1862
- complex representation
 - in optics, 1428
 - of AC, 1310
 - of wave, 1384
- compound microscope, 980
- compound pendulum, 336
- compression ratio
 - in Diesel cycle, 679
 - in Otto cycle, 678
- Compton effect, 1589
 - as elastic collision, 1590
- concave lens
 - diverging action of, 940
 - foci of, 943
 - image formation by, 948
- concurrent forces
 - equilibrium of, 268, 272
 - reduction of, 268
- condenser, 1070
- condition of floatation, 487
- conduction
 - of heat, 731
 - thermal and electrical, 732
- conduction band, 1797, 1799
- conductivity
 - electrical, 1119

conductor, 1114
 conductors
 electrostatics of, 1053
 conservation of energy
 a broad view, 150
 in rotational motion, 233
 conservation principles, 1774
 and symmetry, 1777
 in nuclear reactions, 1760
 constants of motion, 146
 contact electrification, 988
 contact potential, 988
 convection, 732, 740
 natural and forced, 742
 conventional power plant, 674
 convex lens
 converging action of, 940
 focal points of, 942
 image formation by, 947
 Coriolis force, 127, 264
 correlations
 in polarized light, 1494
 cosmic ray, 1594
 Coulomb's law
 in electrostatics, 990
 in friction, 291
 couple, 277
 moment of, 279
 couples
 composition of, 282
 covalent bond, 1780
 critical angle, 900
 critical constants, 722
 critical damping, 353
 critical point, 722
 cross product
 features of, 53
 of vectors, 52
 crystal
 as giant molecule, 1790
 crystals
 energy bands in, 1794
 Curie's law, 1231
 curl
 of vector field, 75
 current
 and current density, 1126
 chemical effect of, 1250
 electrical, 1126
 heating effect of, 1146
 magnetic effect of, 1174
 current density, 1118
 current divider, 1159
 current division, 1159
 current ratio
 in transformer, 1339
 cylinder
 rolling of, 252
 cylindrical capacitor, 1089
 cylindrical current distribution
 magnetic field for, 1207
 Döppler effect, 820

general formula, 823
 uniform motion, 824
 dalton, 20
 damped SHM, 348
 energy dissipation in, 355
 equation of motion for, 348
 general solution, 350
 retardation constant in, 348
 DC circuits
 analysis of, 1160
 DC motor, 1278
 back EMF in, 1281
 de Broglie relations, 1539
 De Morgan's identities, 1869
 decay of charge
 in C - R circuit, 1301
 decay of current
 in L - R circuit, 1295
 decibel, 815
 deformation
 elastic and plastic, 445, 448
 delta connection, 1335
 derived units, 10
 dew point, 728
 and relative humidity, 729
 diamagnetism, 1233
 diathermic enclosure, 592
 diatomic molecules
 heteronuclear and homonuclear, 1789
 dielectric constant, 1069, 1393
 field dependence of, 1397
 dielectric medium
 plane wave in, 1392
 dielectrics
 in capacitors, 1098
 polarization in, 1063
 Diesel cycle, 676, 679
 cut-off ratio in, 679
 efficiency of, 680
 Diesel engine, 676
 differential amplifier, 1853
 diffraction, 1461
 basic theory, 1463
 Fraunhofer and Fresnel, 1469
 diffraction grating, 1486
 diffraction pattern, 1466
 digital circuits, 1861
 combinational, 1861
 sequential, 1861
 digital electronics, 1859
 dimension
 of a physical quantity, 3
 dimensional analysis, 9, 21
 and Stokes' formula, 22
 dimensional homogeneity
 principle of, 21
 dimensions
 related to units, 8
 dipole
 electric, 1012
 magnetic, 1208
 dipole moment, 1014

- of a current distribution, 1211
- dipole source, 782
- dipoles
 - electric and magnetic, 1210
- direction cosine, 50
- disintegration constant, 1756
- dispersion, 872
- dispersion and absorption, 1392
- dispersion relation, 586
- displacement
 - of a particle, 88
 - transformation of, 122
- displacement current, 1352
- divergence
 - of vector field, 75
- division of amplitude
 - in interference, 1445
- division of wave front
 - in interference, 1446
- domain
 - magnetic, 1239
- donor, 1808
- doped semiconductor, 1807
- Doppler effect
 - arbitrary motion, 826
 - in echocardiography, 827
 - relativistic, 1656
- double refraction, 1498
- double slit pattern, 1436
- drag force, 538
- drift velocity, 1115
- dynamic friction, 290
- earth
 - as magnet, 1246
 - potential of, 1101
 - precession of, 260
- echo, 795
- eddy current, 1342
- edge triggering, 1881
- eight-bit arithmetic, 1873
 - addition in, 1874
 - overflow and carry in , 1875
 - subtraction in, 1874
- eikonal surface, 1529
- Einstein's theory, 1578
- elastic collision, 173
- elastic constants, 452
 - for fluids, 466
 - relations between, 463
- elastic deformation
 - mechanism of, 449
- elastic hysteresis
 - and rolling friction, 306
- elastic limit, 445
- elastic waves, 764
 - in anisotropic solid, 867
 - in solids, 864, 866
- elasticity
 - external and internal forces, 424
- electric current
 - and energy pathway, 1139
 - energy transformation in, 1134

electric dipole, 1012, 1015
 force on, 1020
 potential energy of, 1023
 torque on, 1020
 electric displacement
 in dielectric, 1065
 electric field
 at surface of a conductor, 1057
 intensity of, 995
 potential in, 998
 electric field intensity
 unit of, 12
 electric intensity, 995
 electric potential, 998
 and intensity, 999, 1007
 unit of, 12
 electric susceptibility, 1064
 electrical circuits
 Kirchhoff's principles, 1161
 electrical double layer, 1105
 electrical field vectors
 naming of, 1067
 electrical induction, 992
 electrochemical potential, 1803
 electrolysis, 1250
 Faraday's laws of, 1251
 electrolyte, 1104, 1250
 electromagnet, 1244
 electromagnetic field, 1348
 energy density in, 1368
 electromagnetic field tensor, 1663
 electromagnetic induction, 1257
 Faraday's law, 1261, 1350
 electromagnetic interactions, 1775
 electromagnetic spectrum, 1367
 electromagnetic wave, 1347, 1355
 energy flux, 1368
 intensity of, 1371
 phase of, 1364
 phase velocity of, 1363
 plane monochromatic, 1359
 sources of, 1356
 spherical, 1401
 state of polarization, 1378
 transmission of energy, 1357
 wave front for, 1366
 electromagnetism
 in material media, 1354
 electromotive force, 1107
 electron spin, 1695
 electron volt(eV), 19
 electronic configuration, 1698
 and periodic table, 1702
 electronic shells, 1698
 electrons
 free and bound, 1052
 electrostatic shielding, 1075
 electroweak interactions, 1776
 elementary charges, 984
 transfer of, 986
 elementary particles, 1766
 and fields, 1766

- basic interactions, 1772
- classification of, 1767
- quantum numbers of, 1768
- quark structure, 1771
- elliptic orbits
 - and degeneracy, 1691
- elliptic polarization, 1380, 1494
- EMF, 1107
 - as open circuit voltage, 1143
 - in electromagnetic induction, 1144
 - motional, 1264
 - source of, 1142
- emission
 - spontaneous and stimulated, 1501
 - stimulated, 1504
- emission and absorption, 1501
- emission coefficient, 751
- emissivity, 747
- enclosure
 - adiabatic, 591
 - diathermic, 592
- energy
 - and work, 140, 148
 - conservation of, 168
 - kinetic, 94
 - kinetic and potential, 140
 - of charged sphere, 1047
 - of electric dipole, 1023
 - of electric field, 1091
 - of magnetic dipole, 1216
 - of magnetic field, 1304
 - principle of conservation of, 144
 - unit of, 9
- energy and mass
 - equivalence of, 151
- energy bands
 - in crystals, 1794
 - overlapping of, 1796
- energy density
 - acoustic wave, 811
- energy flux
 - electromagnetic wave, 1368
- energy transport
 - velocity of, 1373
- energy-momentum four-vector, 1654
- engine
 - internal combustion, 674
 - reciprocating, 677
- enstrophy, 556
- entropy, 655
 - additivity of, 657
 - as disorder, 663
 - statistical interpretation of, 662
- entropy principle, 655, 663
- equation of motion, 100
 - for fluid, 509
 - for ideal fluid, 510
 - in inertial frame, 124
 - in non-inertial frame, 126
 - in SHM, 320
- equilibrium
 - of concurrent forces, 270

- of forces, 105
 - thermal, 597
 - thermodynamic, 593
- equipotential surfaces, 1031
 - separations between, 1035
- equivalence
 - of heat and work, 613
 - of mass and energy, 151, 1655
- equivalent capacitance
 - parallel connection, 1096
 - series connection, 1095
- equivalent lens, 963
- equivalent resistance, 1153, 1154
- escape velocity, 406
- Euler's equation
 - for ideal fluid, 510
- evaporation, 725
- exact differential, 612
- excess pressure, 767
 - across film, 569
 - amplitude of, 768
 - curved liquid surface, 568
 - in acoustics, 767
 - soap bubble, 570
- exchange force, 1707
- exclusion principle, 1699
- expansion
 - isothermal and adiabatic, 706
- expansion of liquid
 - real and apparent, 693
- expectation value
 - of random variable, 759
- extended fringes
 - in interference, 1447
- Faraday's law
 - electromagnetic induction, 1261
 - in electromagnetism, 1350
- Fermi energy, 1803
- fermions and bosons, 1769
- ferromagnetism, 1237
- field
 - classical and quantum, 1503
 - electromagnetic, 1348, 1353
- field energy
 - electrical, 1093
 - magnetic, 1304
- field of force, 116
 - conservative, 141
- First law of thermodynamics, 611
- flip-flops, 1878
- floatation
 - condition of, 487
 - stable equilibrium in, 490
- fluid
 - buoyancy in, 486
 - equation of motion, 529
 - internal forces in, 474
 - laminar flow of, 529
 - pressure in, 475
 - steady flow of, 506
 - stream line in, 504
 - thrust of, 480

- transmission of pressure in, 499
- fluid flow
 - irrotational, 507
 - stability of, 551
- fluids
 - elastic constants for, 466
 - non-Newtonian, 530
- flux
 - magnetic, 1259
 - of electric field intensity, 1036
 - of gravitational intensity, 383
- focal length
 - spherical lens, 945
 - spherical mirror, 920
 - spherical surface, 934
- force
 - central, 158
 - centrifugal, 264
 - conservative, 142
 - Coriolis, 264
 - electrostatic, 990
 - impulse of, 188
 - impulsive, 191
 - internal and external, 156
 - line of action of, 102
 - moment about a point, 217
 - moment about an axis, 220
 - non-central, 158
 - non-conservative, 142
 - point of application of, 103
 - qualitative idea of, 98
 - tidal, 414
 - unit of, 9
- force and couple
 - reduction of, 283
- force and torque
 - on magnetic dipole, 1215
- force four-vector, 1660
- forced convection, 742
- forced SHM, 357
 - energy exchange in, 361
 - general solution for, 358
 - resonance in, 361
 - steady state solution, 359
 - transient solution, 359
- forces
 - composition of, 104
 - in equilibrium, 105
 - resultant of, 104
- forces on rigid body
 - reduction of, 275
- forward bias, 1818
- four-momentum, 1654
- four-vectors, 1641
 - contravariant and covariant, 1643
 - inner product of, 1648
 - space-like and time-like, 1648
- four-velocity, 1654
- frame of reference, 81, 1596
 - inertial, 97, 124
 - non-inertial, 117
 - rotating, 263

frames of reference
transformations between, 117

Fraunhofer diffraction
phase in, 1479

Fraunhofer pattern
double slit, 1481
grating, 1486
single slit, 1471

free expansion, 651

free particle
in quantum theory, 1536
wave-like features of, 1538

frequency
unit of, 11

frequency and phase, 821

Fresnel formulae, 898

friction, 288
Amonton's law in, 290
and wear, 302
Coulomb's law in, 291
dynamic, 290
elastic and plastic deformations in, 301
mechanism of, 296
rolling, 305
static, 289
static and dynamic, 288
stick-and-slip in, 301
wet, 304

fringes
of equal inclination, 1449
of equal thickness, 1451

Galilean principle of equivalence, 126, 1597

Galilean transformation, 123, 1603

gamma decay, 1754

gas
kinetic theory of, 623
thermal expansion of, 698

gas constant, 626

Gauss' principle
applications of, 386, 1043
derivation of, 1042
for dielectric, 1067
in electrostatics, 1036
in gravitation, 385

Gaussian optics, 971

general principle of equivalence, 1667

general relativity
electromagnetic field in, 1675
Schwarzschild solution in, 1676

general theory of relativity, 1667

generalrelativity
Einstein's equations in, 1672

generator
AC and DC, 1268

geomagnetic dynamo, 1248

geomagnetism, 1246

geometrical optics, 1413

geometrical wave front, 789

geosynchronous orbit, 397

global warming, 755, 756

gluon, 1776

gradient

- of scalar field, 75
- gravitation
 - broader view of, 420
 - Gauss' principle in, 385
 - Newton's law of, 367
- gravitational constant, 368
 - unit of, 15
- gravitational field
 - equations of motion in, 1674
- gravitational intensity, 372
 - due to spherical mass, 386
 - flux of, 383
- gravitational potential, 376
- gravitational time dilatation, 1680
- graviton, 1776
- gravity waves, 586
- greenhouse effect, 754
- group velocity, 586, 848, 1404
- growth of charge
 - in C - R circuit, 1299
- hadrons, 1767
- half cell, 1104
 - standard, 1105
 - standard hydrogen, 1105
- half cell potential, 1106
- half life, 1757
- half-adder, 1878
- harmonic oscillator
 - in black body radiation, 1562
 - in quantum theory, 1550
 - in thermal equilibrium, 1564
- Hartree theory, 1708
- heat, 592
 - direction of flow, 602
 - generated in friction, 292
 - quantitative definition of, 610
 - stationary flow of, 737
 - transmission of, 731
 - unit of, 613
- heat and work
 - equivalence of, 613
- heat engine, 664
 - efficiency of, 666, 667
 - reversible, 665
- heat flow
 - non-stationary, 737
- heat sink, 603
- heat source, 603
- heavy top
 - precession of, 259
- hemisphere
 - center of mass of, 164
- hertz, 11
- Higgs boson, 1777
- Higgs field, 1777
- hole, 1802
- hologram, 1512
 - of a point object, 1514
- holography, 1511
- human eye, 973
- hydrogen atom
 - Bohr's theory of, 1566

- hydrogen bond, 1783
- hydrogen spectrum, 1567
- hysteresis, 1227, 1242
- ideal fluid, 507
 - equation of motion, 509, 510
- ideal gas, 626
 - adiabatic process in, 638
 - equation of state of, 626, 636
 - internal energy of, 636
 - isothermal process in, 638
 - velocity of sound in, 807
 - work performed by, 639
- image formation
 - aberration in, 884
 - by multiple reflection, 887
 - by reflection, 885
 - by refraction, 889
 - by spherical surface, 932
 - off-axis point, 952
 - spherical lens, 945
 - spherical mirror, 920
- image imperfection, 972
- impedance
 - complex, 1316
 - in AC, 1316, 1318
 - series and parallel, 1321
- impedance matching, 1339
- impulse of a force, 188
- impulse of a torque, 221
- impulsive force, 191
- impulsive torque, 221
- incoherent mixture, 1405
- indistinguishability, 1700
- induction
 - electromagnetic, 1257
- inelastic collision, 173
- inertial force, 127
- inertial frame, 97, 124, 1600
 - equation of motion in, 124
- inexact differential, 612
- inhomogeneous wave, 786
- intensity
 - gravitational, 372
 - inverse square law of, 814
 - law of, 879
 - maxima and minima, 1424, 1432
 - of electromagnetic wave, 1371
 - of sound, 812
- interference
 - basic idea, 836
 - conditions for, 1426
 - extended fringes in, 1447
 - fringes of equal inclination, 1449
 - of scalar waves, 1427
- internal energy, 608
 - of ideal gas, 636
 - significance of, 609
- internal modes
 - energy of, 171
 - in collisions, 171
- intrinsic parity, 1770
- intrinsic semiconductor, 1805

- inverse square field
 - bounded motion in, 408
 - circular orbit in, 412
 - equation of motion in, 408
 - unbounded motion in, 408
- inverse square law
 - of intensity, 814
- ion, 1688
- ionic bond, 1779
- irrotational flow, 507
- isothermal curve, 599
- isothermal process, 604
 - work in, 639
- Joule's law, 1146
- junction diode, 1812
 - AC resistance of, 1820
 - as rectifier, 1823
 - at equilibrium, 1814
 - barrier height, 1817
 - biased, 1818
 - characteristic, 1819
 - DC resistance of, 1820
- Jurin's law, 572
- kelvin
 - definition of, 7
- Kelvin equation, 580
- kelvin scale of temperature, 601
- Kepler's laws, 409
 - of planetary motion, 409
- kilogram
 - definition of, 7
- kinematic viscosity, 526
- kinetic energy, 137
 - in rotation, 233
 - of a particle, 94
 - related to work, 140
- kinetic theory, 623
- Kirchhoff's principle
 - in AC, 1324
 - in radiation, 750
- Kirchhoff's principles
 - in electricity, 1161
- Kolmogorov scale
 - in turbulence, 557
- laminar flow, 507, 529
- Laplace formula
 - for pressure difference, 569
- laser
 - as coherent source, 1500
- laser diode, 1511, 1831
 - population inversion in, 1832
- latent heat, 714
 - of evaporation, 715
 - of fusion, 714
- lead-acid accumulator, 1113
- Leclanche cell, 1112
- LED, 1830
- Lenz's law, 1263
- lepton number, 1770
- leptons, 1767
- lift force, 515

- light
 - polarized and unpolarized, 1492
- light cone, 1636
- light emitting diode, 1830
- like parallel forces
 - reduction of, 275
- limiting friction
 - and yield stress, 298
 - force of, 289, 298
- line integral
 - closed, 78
 - of vector field, 78
- linear expansion
 - coefficient of, 686
- linear polarization, 1380
- linear vector space, 31, 40
 - dimension of, 32
- lines of force
 - in electric field, 1027
 - magnetic, 1188, 1350
 - magnetic and electric, 1217
- liquid
 - surface energy of, 559
 - surface tension of, 560
 - temporary binding in, 1791
- liquid drop model, 1731
- liquid surface
 - pressure difference across, 568
- logic families, 1867
- logic gate
 - EXOR, 1868
 - NAND, 1868
 - NOR, 1868
- logic gates
 - elementary, 1862
- Lorentz contraction, 1616
- Lorentz force, 1183
- Lorentz transformation, 1602
 - general form, 1609
- loudness, 863
- lubrication, 304
- Mach number, 830
- macrostate and microstate, 623
- Madelung rule, 1710
- magnetic dipole, 1208
 - energy of, 1216
 - force on, 1215
 - torque on, 1215
- magnetic domain, 1239
- magnetic field
 - of circular wire, 1194
 - of current, 1180
 - of solenoid, 1197
 - of straight wire, 1191
- magnetic field energy, 1304
- magnetic field strength
 - unit of, 12
- magnetic field vectors
 - naming of, 12, 1185
- magnetic flux, 1259
- magnetic flux density
 - unit of, 12

magnetic force
 between currents, 1177
 on current, 1180
 on moving charge, 1183
 magnetic intensity
 due to current, 1185
 magnetic lines of force, 1217
 magnetic susceptibility, 1224
 magnetization
 spontaneous, 1238
 magnetization curve, 1241
 magnetization vector, 1223
 magnification
 angular, 957
 longitudinal, 956
 transverse, 956
 majority carrier, 1811
 mass and energy
 equivalence of, 151
 mass formula, 1740
 mass number, 1725
 mass-energy equivalence, 1655, 1733
 master-slave flip-flop, 1886
 materials
 magnetic properties of, 1222
 matter
 phases of, 473
 states of, 472
 Maxwell's equations, 1348
 Maxwell's velocity distribution formula, 648
 mean free path, 643
 mean life, 1756
 mean speed, 645
 mechanical equivalent of heat, 613
 melting point, 714
 mercury battery, 1113
 mesons, 1767
 metacenter, 491
 metacentric height, 493
 metamaterials, 1400
 meter
 definition of, 7
 metric tensor, 1644
 micelles, 588
 micro-emulsion, 589
 microscope, 980
 microstate, 623
 Mie scattering, 1524
 Minkowski geometry, 1636
 minority carrier, 1811
 mixed states
 in quantum theory, 1559
 mixed strain, 435
 mixed stress, 442
 mobility of carriers, 1118
 modulus of rigidity, 459
 mol
 definition of, 7
 mole number, 625
 molecular bonding, 1779
 molecular collisions, 633, 641
 molecular excitations, 1785

molecule
 as unit of matter, 1686
 binding of, 1778
 electronic states of, 1786
 random motion of, 641
 rotational states of, 1787
 stationary states of, 1785
 vibrational states of, 1786
 moment of a force, 217
 about a point, 217
 about an axis, 220
 moment of inertia, 226
 additivity of, 238
 annular ring, 242
 calculation of, 234
 cylindrical shell, 238
 of cylinder, 242
 of disc, 240
 of ring, 238
 of sphere, 244
 momentum
 of a particle, 93
 conservation of, 167
 of center of mass, 162
 momentum uncertainty, 1541
 and position uncertainty, 1543
 monochromatic wave, 1359
 monopole source, 780
 moon
 orbit of, 417
 Mosley's law, 1721
 most probable speed, 645
 motion
 bounded and unbounded, 405
 rectilinear, 106
 uniform circular, 203
 motional EMF, 1264
 moving mass, 151
 multiplication of force
 principle of, 501
 musical instruments, 862
 mutual capacitance, 1083
 mutual inductance, 1288
 natural convection, 742
 Navier-Stokes equation, 529
 negative feedback, 1849, 1856
 negative refractive index, 881, 1400
 neutral point, 1027
 neutrino, 1751
 types of, 1751
 newton, 9
 Newton's first law, 97
 Newton's formula
 for a fluid, 524
 thin lens, 951
 Newton's law of cooling, 752
 Newton's law of gravitation, 367
 Newton's rings, 1453
 Newton's second law, 99
 Newton's third law, 157
 Newtonian fluid
 turbulent motion of, 529

- no-slip condition, 516, 534
- non-concurrent forces
 - reduction of, 269
- non-coplanar forces
 - reduction of, 285
- non-inertial frame, 117
 - equation of motion in, 126
 - inertial force in, 127
- non-Newtonian fluids, 530
- non-ohmic elements, 1792
- non-paraxial rays
 - in image formation, 891
- non-reflective coating, 1460
- non-SI units, 19
- normal
 - for space curve, 209
- normal modes, 843
 - in a crystal, 865
 - of a stretched diaphragm, 860
- normal reaction, 300
- NOT gate
 - with transistor, 1866
- nuclear charge
 - and binding energy, 1740
- nuclear excitation, 1744
 - collective, 1746
 - single-particle, 1745
- nuclear fission, 1762
- nuclear force
 - saturation property, 1729
- nuclear fusion, 1764
- nuclear mass, 1733
 - and binding energy, 1735
 - unit of, 1734
- nuclear mass formula, 1740
- nuclear radius, 1731
- nuclear reactions, 1758
 - conservation principles, 1760
 - energy balance in, 1760
- nuclear stability, 1736
- nucleon, 1687
 - internal characteristics, 1725
- nucleons
 - interaction of, 1727
- nucleus
 - as liquid drop, 1731
 - binding energy of, 1735
- null vector, 35
- object-image relation
 - Newton's formula, 951
 - spherical lens, 947
 - spherical mirror, 923
- octets, 1768
- off-axis point
 - image formation, 927, 952
- Ohm's law, 1130
- oil
 - calming effect of, 581
- oil over water, 581
- operational amplifier, 1853
- optical anisotropy, 1498
- optical instruments, 975

optical path, 1434
 optical resonator, 1509
 OR and AND gates
 with diodes, 1864
 orbit of moon, 417
 orbital, 1574, 1695, 1781
 orthonormal triad, 44
 left handed, 45
 right handed, 45
 oscillations
 of physical quantities, 761
 simple harmonic, 761
 Otto cycle, 676, 677
 efficiency of, 678
 overdamped SHM, 353
 p-n diode, 1812
 parallelogram rule, 36
 paramagnetism, 1228
 paraxial ray, 911, 914
 paraxial rays, 890
 partial pressure, 650
 partially coherent light, 1439
 particle
 acceleration of, 95
 displacement of, 88
 equation of motion of, 100
 kinetic energy of, 94
 momentum of, 93
 planar motion of, 111
 velocity of, 89
 partition function, 662
 Pascal's law, 499
 path difference, 1434
 Pauli's principle, 1699
 Peltier effect, 1254
 pendulum
 compound, 336
 simple, 318, 341
 periodic table, 1702
 permeability, 1224
 of free space, 16
 permittivity
 of free space, 16
 persistence time, 796
 petrol engine, 676
 phase
 in AC, 1309
 in Fraunhofer diffraction, 1479
 in progressive wave, 772
 in sound wave, 769
 of electromagnetic wave, 1364
 phase difference, 1433
 phase transition, 712
 phase velocity, 586, 775
 in dielectric medium, 1396
 of plane wave, 772
 of spherical wave, 781
 phasor, 1312
 photo-ionization, 1583
 photoelectric effect, 1578
 photoelectric emission
 features of, 1579

- role of photons, 1581
- photon, 1563
 - state of, 1502
- physical quantities
 - as scalars and vectors, 27
 - as tensors, 29
 - dimensionless, 9
 - dimensions of, 3
 - fundamental units of, 4
- pin-holes and slits, 1419
- pitch, 863
- planar motion
 - of a particle, 111
- Planck distribution formula, 1561
- Planck's constant
 - unit of, 16
- plane polar co-ordinates, 212
- plane progressive wave, 770
- plane wave, 774
 - field variables in, 1360
 - in conducting medium, 1397
 - in dielectric medium, 1392
 - reflection and refraction of, 1386
 - wave vector of, 1382
- plane waves
 - superposition of, 1421
- planetary motion
 - Kepler's laws of, 409
- plastic deformation
 - mechanism of, 450
- Poiseuille's flow, 532
- Poisson's ratio, 455
- polar vector, 54, 219
- polarization
 - by scattering, 1500
 - of electromagnetic wave, 1378
- polarized light
 - production of, 1499
- polaroid, 1499
- population inversion, 1507
- position vector, 70
 - time dependence of, 87
- potential
 - at infinity, 1000
 - electrical, 998
 - of point charge, 1003
- potential energy, 141
 - gravitational, 136
 - in rectilinear motion, 134
 - in rotation, 233
- potential flow, 514
- power, 133
 - in AC circuit, 1326
 - unit of, 10
- power factor, 1328
- power plant
 - conventional, 674
- Poynting vector, 1370
- Prandtl number, 27
- precession, 257
 - of earth's axis, 260
 - of heavy top, 259

- torque-free, 262
- under an applied torque, 257
- pressure
 - and momentum transfer, 632
 - atmospheric, 484
 - in a liquid, 477
 - in kinetic theory, 627
 - of an ideal gas, 632
 - unit of, 10
- principle of conservation of energy, 144
 - in rectilinear motion, 139
- principle of equivalence, 126
- principle of similarity, 24
- principle of superposition
 - in AC circuits, 1325
 - in DC circuits, 1164
 - in deformation, 461
 - in electromagnetism, 1357
 - in electrostatics, 993
 - in gravitation, 369
 - in magnetism, 1185
 - in quantum theory, 1559
 - in wave motion, 777
- principle of uncertainty, 1541
- prism, 902
 - deviation by, 904
 - minimum deviation by, 907
 - refraction by, 903
- probability
 - of microstates, 662
- probability distribution, 757
- moments of, 760
- process
 - adiabatic, 605
 - irreversible, 651
 - isothermal, 604
 - reversible, 653
 - thermodynamic, 594
- processes
 - elastic and inelastic, 171
- progressive wave, 770
 - phase in, 772
 - phase velocity of, 772
- projectile
 - motion of, 115, 419
- pseudo-force, 127
- Q-point, 1847
 - stability of, 1847
- Q-value, 1761
- quadrupole source, 784
- quality of sound, 863
- quantized energy, 1551
- quantum and classical
 - optical analogy, 1532
- quantum number
 - magnetic, 1694
- quantum numbers
 - of elementary particles, 1768
- quantum optics, 1413
- quantum theory
 - free particle in, 1536
 - mixed states in, 1559

- of particles and fields, 1593
- random variables in, 1535
- simple harmonic oscillator, 1550
- state of a system, 1534
- superposed states in, 1558
- time evolution in, 1556
- quarks, 1771
- quasi-monochromatic light, 1440
- quasi-static process, 596
 - work performed in, 613
- radiation
 - energy exchange in, 748
- radiation pressure, 1376
- radioactive decay, 1747
- radioactive decay law, 1755
- radius of gyration, 237
- raindrops
 - formation of, 578
- Raman scattering, 1525
- random motion
 - of molecules, 641
- random variable, 757
 - expectation value of, 759
 - in quantum theory, 1535
- Rankine cycle, 674
- ray and wave normal, 1383
- ray optics, 1413
 - basic principles, 871
 - related to wave optics, 1526
 - sign convention in, 893, 916
- ray path, 871
- Rayleigh scattering, 1517
- Rayleigh-Plateau instability, 583
- rays
 - image formation by, 882
- reactance
 - in AC circuit, 1319
 - of an inductor, 1320
 - of capacitor, 1320
- real gas
 - behavior of, 637
- reciprocating engines, 677
- rectifier, 1823
- rectilinear motion, 106
 - potential energy in, 134
 - principle of conservation of energy in, 139
- red shift, 1680
- reduced mass, 408
- reduction
 - of a system of forces, 268
 - of co-planar forces, 284
 - of force and couple, 283
 - of non-co-planar forces, 285
 - of system of forces, 275
- reflection
 - laws of, 877
 - total internal, 897
- reflection and refraction, 789
 - of electromagnetic wave, 1386
- refraction
 - at spherical surface, 930

laws of, 877
 refractive index, 872
 negative, 881
 refrigerants, 683
 refrigeration, 681
 refrigerator, 652, 681
 work ratio of, 682
 relative humidity, 727
 relative permeability, 1203
 relative permittivity, 1069
 relative velocity, 119
 relativistic aberration, 1625
 Relativistic Döppler effect, 1656
 relativistic energy, 1651
 relativistic mass, 1651
 relativistic momentum, 1651
 relativity
 general theory, 1667
 special theory, 1601
 relativity of simultaneity, 1615
 residual magnetism, 1242
 resistance, 1129
 equivalent, 1153, 1154
 resistances
 in series and parallel, 1151
 resistivity, 1119
 temperature dependence of, 1133
 resolving power, 1489
 resonance
 in AC, 1317
 in forced SHM, 361
 resonant cavity, 1509
 rest energy, 1655
 rest mass, 151, 1651
 restoring force
 in SHM, 319
 in spring, 345
 restoring torque
 in compound pendulum, 336
 resultant
 of forces, 104
 reverberation, 797
 time, 799
 reverse bias, 1818
 reverse saturation current, 1820
 Reynolds number, 25, 535
 right hand rule, 59
 rigid body
 motion of, 248
 rotational motion, 248
 translational motion, 248
 ripples, 586
 RMS speed, 632
 RMS value
 of AC voltage, 1309
 of AC current, 1309
 RMS velocity, 632
 rolling
 of cylinder, 252
 of sphere, 254
 of wheel, 254
 rolling friction, 305

- Coulomb's law in, 309
- elastic hysteresis in, 306
- plastic deformation in, 307
- plastic hysteresis in, 307
- rolling motion, 251
- rotating frame, 263
 - pseudo-force in, 263
- rotating magnetic field, 1274
- rotational kinetic energy, 233
- rotational motion
 - about an axis, 225
 - conservation of energy in, 233
 - work done in, 231
- rotational potential energy, 233
- Rydberg constant, 1569
- saturated air, 722
- saturated vapor, 718
- saturation pressure, 719, 723
 - over curved surface, 580
- saturation vapor pressure, 719
- scalar field, 72
 - derivative of, 74
 - gradient of, 75
 - volume integral of, 76
- scalar product, 42
 - features of, 43
- scalar triple product, 64
- scalars, 27, 28
 - as real numbers, 33
 - related to a vector space, 32
- scattering
 - Mie, 1524
 - Raman, 1525
 - Rayleigh, 1517
- Schrödinger equation, 1557
- Schwarzschild solution, 1676
 - Newtonian limit of, 1679
- screening
 - of nuclear charge, 1703
- second
 - definition of, 6
- secondary wave, 1465
- Seebeck effect, 1253
- selection rule, 1508, 1788
 - atomic, 1724
 - for vibrational transitions, 1789
- self inductance
 - of long solenoid, 1285
- self-capacitance, 1083
- self-inductance, 1284
 - of toroidal solenoid, 1286
- semiconductors, 1794
 - direct bandgap, 1830
 - doped, 1807
 - intrinsic, 1805
- sequential circuits, 1878
- sequential logic, 1862
- shear, 432
- shear modulus, 459
- shear stress, 440
- shear viscosity, 528
- shear wave, 867

shearing strain, 432
 shock front, 831
 shock wave, 830
 SI scale
 of temperature, 602
 SI system
 base units in, 4, 6
 derived units in, 4
 sign convention, 877, 893, 913, 916
 for refractive index, 918
 similarity
 principle of, 24
 simple harmonic motion, 319
 equation of motion in, 320
 amplitude of, 326
 angular, 336, 337
 angular frequency of, 322
 conservation of energy, 331
 critical damping in, 350
 critically damped, 353
 damped, 348
 force constant in, 319
 forced, 357
 frequency of, 325
 general and particular solutions, 322
 graphical representation of, 328
 initial conditions in, 323
 isochronicity in, 344
 kinetic energy in, 331
 overdamped, 350, 353
 periodicity of, 324
 phase in, 325
 potential energy in, 330
 time period of, 325
 underdamped, 350, 351
 velocity in, 324
 simple harmonic oscillator
 quantum, 1550
 simple pendulum, 318, 341
 sine rule
 for equilibrium, 271
 single-electron states
 and space quantization, 1694
 siphon, 516, 517
 speed of efflux in, 520
 velocity of efflux, 520
 Snell's law, 878
 solenoid
 magnetic field of, 1197
 self-inductance of, 1285
 solid angle, 1039
 solids
 cohesion of, 1790
 thermal expansion of, 686
 solubilization, 589
 sonic boom, 831
 sound
 as a pressure wave, 764
 intensity of, 812
 reverberation of, 797
 sound wave
 velocity of, 804

- sound waves
 - and rays, 788
 - coherence of, 834
 - diffraction of, 791
 - energy density in, 810
 - excess pressure in, 767
 - in one dimension, 767
 - interference of, 832
 - reflection and refraction of, 787
 - scattering of, 791
- source of sound
 - dipole, 782
 - monopole, 780
 - quadrupole, 784
- space curve
 - local description of, 209
- space cuve
 - motion along, 209
- space quantization, 1694
- space-time diagram, 1626
 - invariant regions, 1630
- space-time geometry, 1633
- space-time interval, 1606
- space-time separation, 1606
 - invariance of, 1612
 - light-like, 1633
 - space-like, 1632
 - time-like, 1632
- special theory of relativity, 1601
- specific binding energy, 1736
- specific heat
 - at constant pressure, 705
 - at constant volume, 704
- specific heats, 702
 - of ideal gas, 703
 - ratio of, 708
- speed
 - of a particle, 90
 - as magnitude of velocity, 91
 - root mean squared, 632
- sphere
 - rolling of, 254
- spherical aberration, 970
 - longitudinal, 970
 - transverse, 970
- spherical lens, 939
 - convex and concave, 939
 - focal lengths of, 945
 - image formation by, 945
 - object-image relation, 947
 - radii of curvature, 944
 - thin, 941
- spherical mirror
 - center of curvature, 909
 - concave and convex, 909
 - focal length, 920
 - image formation, 920
 - object-image relation, 923
 - off-axis point, 927
- spherical surface
 - image formation by, 932
 - refraction at, 913, 930

spherical wave, 780

spin

- of electron, 1695

spin quantum number, 1696

- magnetic, 1696

spontaneous emission, 1501, 1724

spring

- energy of deformation, 345
- restoring force in, 345

stability

- and binding energy, 1736
- of floating body, 490
- of fluid flow, 551

standing wave, 838

- as superposition, 841
- fundamental frequency, 840
- in air column, 838
- nodes and antinodes, 844
- normal modes in, 843

star connection, 1331, 1332

state

- quantum description of, 1534

state diagram

- thermodynamic, 595

state functions, 594

state of motion, 102

state variables, 594

static friction, 289

stationary wave, 838, 1408

steady flow

- of a fluid, 506

Stefan's constant, 746

- unit of, 16

Stefan's law, 746

stick-and-slip

- in friction, 301

stimulated emission, 1501, 1724

- and coherence, 1505

Stokes' formula, 545

stopping potential, 1587

storage cell, 1113

straight wire

- magnetic field of, 1191

strain, 425

- bulk or volume, 431
- definition of, 427
- homogeneous, 430
- in acoustic wave, 803
- mixed, 435
- principal axes of, 435
- principal components of, 435
- shear, 432
- tensile, 430

strain energy, 467

- density, 468, 470

strangeness, 1770

stream line

- in a fluid, 504

stress, 426

- as a surface force, 437
- bulk, 441
- description of, 439

- in a deformable body, 439
 - mixed, 442
 - principal axes of, 442
 - principal components of, 442
 - shearing, 440
 - tensile, 439
 - unit of, 10
- stress-strain curve, 444
- stress-strain relations, 452
- stretched diaphragm
 - vibrations of, 858
- stretched string
 - fundamental mode, 856
 - harmonics in, 857
 - vibrations of, 852
 - wave equation for, 853
- strings and diaphragms
 - vibrations of, 851
- strong interaction, 1726
- successive disintegrations, 1757
- superposition
 - of plane waves, 1421
 - principle of, 369, 461, 777, 993, 1164, 1185, 1325, 1357, 1559
- supersonic object, 830
- surface charge
 - on conductor, 1054
- surface charge density, 1055
- surface energy, 558
- surface force
 - on charged conductor, 1059
- surface integral
 - closed, 77
 - of a vector field, 77
- surface tension, 560
 - and pressure difference, 569
 - as lateral force, 563
 - unit of, 11
- surface waves, 585
 - group velocity, 586
 - phase velocity, 586
 - velocity of, 585
- surfacial expansion
 - coefficient of, 687
- surfactants, 587
 - as detergents, 588
 - effect on surface tension, 588
 - molecular structure of, 587
- susceptibility
 - magnetic, 1224
- SVP, 719
 - over curved surface, 580
- SVP over curved surface, 580
- symmetry
 - and conservation principles, 1777
- synchronous motor, 1276
- system of forces
 - reduction of, 268, 275
- system of particles
 - angular motion of, 222
 - center of mass of, 159
 - mechanics of, 156

temperature

- absolute scale of, 601, 672
- as a thermodynamic variable, 598
- Celsius and Fahrenheit scales, 600
- empirical scales of, 599
- gas scale of, 635
- kinetic interpretation of, 632
- scales of, 673
- SI scale of, 601, 602, 674

tensile strain

- homogeneous, 430

tensile stress, 439

tensor field, 1670

tensors, 29, 1642

theorem of parallel axes, 236

theorem of perpendicular axis, 234

thermal capacity, 700

- at constant pressure, 701
- at constant volume, 701

thermal conductivity, 733

thermal diffusivity, 737

thermal equilibrium, 597

thermal expansion, 684

- of liquids, 693
- coefficients of, 686
- mechanisms of, 684
- of gases, 698
- of solids, 686

thermal radiation, 742

- Kirchhoff's principle in, 750
- Stefan's law of, 746

thermal reservoir, 603

thermal stress, 691

thermodynamic equilibrium, 593

thermodynamic process, 594

thermodynamic scale of temperature, 672

thermodynamic systems, 590

thermodynamic variables, 594

- intensive and extensive, 622

thermodynamics

- first law of, 607, 611
- second law of, 655
- zeroth law of, 597

thermoelectric effects, 1252

thermonuclear reactions, 1766

thin film interference, 1443

thin lens

- image formation by, 945

thin lenses

- combination of, 961
- in contact, 966

three-phase supply, 1329

three-phase transformer, 1341

threshold frequency, 1587

thrust

- of a fluid, 480

tidal force, 414

time

- relativity of, 85

time constant

- of C - R circuit, 1301
- of L - R circuit, 1294

time dilatation, 1619
 time evolution
 in quantum theory, 1556
 toroidal solenoid
 self-inductance of, 1286
 torque, 218
 impulse of, 221
 impulsive, 221
 total internal reflection, 897, 1391
 trajectory of a particle, 88
 transformation
 of acceleration, 123
 of displacement, 121, 122
 of time, 121
 transformation of vectors, 60
 transformer, 1335
 current ratio in, 1339
 energy loss in, 1340
 voltage ratio in, 1337
 transistor, 1833
 α and β , 1842
 characteristics, 1838
 currents and voltages in, 1836
 DC bias, 1847
 doping in, 1834
 forward current gain, 1844
 in active mode, 1836
 in NOT gate, 1866
 in voltage amplification, 1846
 input impedance, 1845
 n-p-n, 1834
 p-n-p, 1834
 saturation and cut-off, 1838
 two-diode model, 1835
 transistor amplifier
 voltage gain, 1850
 transition temperature, 714
 dependence on pressure, 717
 transverse vibrations
 of stretched string, 852
 triangle rule
 for equilibrium, 270
 triple point, 719
 triple point of water, 7
 truth table, 1862
 adder, 1877
 elementary gates, 1863
 EXOR gate, 1868
 universal gates, 1868
 tubes of force, 1033
 tuning fork
 as quadrupole source, 784
 sound wave from, 765
 turbulence, 507
 and viscosity, 529
 and vorticity, 551
 Kolmogorov scale in, 557
 turbulent flow, 507
 twin paradox, 1622
 two-diode model, 1835
 u-v relation
 spherical lens, 947

- spherical mirror, 923
- ultrasonic waves, 816
- uncertainty principle, 1541
 - and harmonic oscillator, 1553
- underdamped SHM, 351
- uniform circular motion, 203
- uniqueness theorem, 1074
- unit
 - coefficient of viscosity, 11
 - electric charge, 11
 - electric field intensity, 12
 - electric potential, 12
 - energy, 9
 - frequency, 11
 - magnetic field strength, 12
 - magnetic flux density, 12
 - power, 10
 - pressure, 10
 - stress, 10
 - surface tension, 11
- unit vector, 40
 - direction cosines of, 50
- units
 - of physical quantities, 1
 - basic and derived, 4
 - cgs system of, 18
 - conversion of, 19
 - derived, 10
 - prefixes in, 18
 - SI system of, 2
- universal gas constant
 - unit of, 15
- unlike parallel forces
 - reduction of, 277
- unpolarized wave, 1380
- valence band, 1797, 1798
- varying currents, 1290
 - analysis of, 1296
- vector
 - broader definition of, 31
 - Cartesian components of, 45
 - line integral of, 78
 - magnitude of, 31, 34
 - multiplication with a scalar, 37
 - null, 35
 - scalar component of, 56
 - two dimensional, 47
 - vectorial component of, 56
- vector addition, 35, 36
 - features of, 38
- vector field, 73
 - curl of, 75
 - derivative of, 74
 - divergence of, 75
 - line integral of, 78
 - surface integral of, 77
- vector function
 - derivative of, 68
 - of scalar variable, 67
- vector product, 52
 - features of, 53
- vector triple product, 66

- vector variable
 - scalar function of, 70
- vectorial area, 55
- vectors, 27, 28, 31
 - addition of, 35, 36
 - as directed line segments, 31
 - equality of, 33
 - linearity of, 32
 - operations with, 35
 - orthonormal triad, 44
 - scalar product of, 42
 - transformation of, 60
 - triple products of, 64
 - vector product of, 52
- velocity
 - of a particle, 89
 - relative, 119
 - transformation of, 117
- velocity four-vector, 1649
- velocity of light, 1601, 1614
- velocity of sound
 - Laplace's formula, 808
 - Newton's formula, 809
- velocity of sound in air, 809
- velocity potential, 507, 514
- velocity transformation, 1622
- viscosity, 523
 - and turbulence, 529
 - as internal friction, 523
 - as momentum transport, 527
 - coefficient of, 525
 - kinematic and dynamic, 526
 - Newton's formula for, 524
 - no-slip condition, 516
 - origin of, 536
 - temperature coefficient of, 538
 - temperature dependence of, 526
- viscous drag
 - Stokes' formula for, 545
- viscous friction, 304
- voltage amplifier, 1846
- voltage divider, 1157
- voltage division, 1157
- voltage gain, 1850
- voltage ratio
 - in transformer, 1337
- voltage regulation, 1827
- volume expansion, 687
- volume integral
 - of a scalar field, 76
- volume strain, 431
- volume stress, 441
- vorticity
 - and turbulence, 551
 - in fluid flow, 507
- walking on water, 582
- water
 - anomalous expansion of, 696
- wave, 762
 - acoustic, 764
 - as transmission of oscillation , 763
 - complex representation, 1384

- differential equation for, 778
 - diffraction and scattering, 791
 - electromagnetic, 1355
 - in three dimensions, 774
 - inhomogeneous, 786
 - monochromatic, 773, 775
 - phase velocity of, 775
 - plane progressive, 770, 775
 - principle of superposition, 777
 - reflection and refraction, 789
 - spherical, 780
 - ultrasonic, 816
- wave equation, 778
 - boundary conditions in, 780
 - for stretched diaphragm, 859
 - in three dimensions, 779
- wave front
 - and ray, 789
 - for a plane wave, 774
 - in electromagnetic wave, 1366
- wave function, 764
 - in quantum theory, 1539
- wave normal, 775
- wave normal and ray, 1383
- wave optics, 1413
- wave packet, 848, 1404
 - group velocity of, 850
- wave vector, 1382
- wave-like features
 - of free particle, 1538
- waves
 - coherent, 1382
 - coherent and incoherent, 1405
 - electromagnetic, 1347
 - partially coherent, 1382
 - spreading and bending of, 1417
- weak interaction, 1751
- wear
 - in friction, 302
- weight
 - as force of reaction, 400
 - as gravitational pull, 400
- weightlessness, 401
- wet friction, 304
- wetting of a surface
 - by a liquid, 566
- Wheatstone bridge, 1166
- wheel
 - couple-driven, 310
 - force-driven, 313
 - motion of, 309
 - rolling of, 254
 - under external driving, 309
- Wien bridge oscillator, 1858
- work, 591
 - and energy, 140
 - adiabatic, 607
 - done by a force, 130
 - expression for, 130
 - in adiabatic process, 640
 - in isothermal process, 639
 - in rectilinear motion, 134

- in rotational motion, 231
- work and heat
 - in surface energy, 560
- world line, 1626
- wrench, 285
- X-ray spectrum, 1717
 - characteristic, 1717
 - continuous, 1717
- Young formula
 - for angle of contact, 566
- Young's modulus, 453
- Young's pattern, 1430
 - with unpolarized light, 1439
- Zener diode, 1827
 - as regulator, 1827
 - characteristic of, 1827
 - doping in, 1827